On the optimality of regularity in mixing Markovian decision rules for MDP control

van der Laan, D.A.

2010

TI 2010-036/4
Tinbergen Institute Discussion Paper

On the Optimality of Regularity in Mixing Markovian Decision Rules for MDP Control

Dinard van der Laan

Department of Econometrics and OR, VU University Amsterdam, and Tinbergen Institute.

Tinbergen Institute

The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam, and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Roetersstraat 31
1018 WB Amsterdam
The Netherlands
Tel.: +31(0)20 551 3500
Fax: +31(0)20 551 3555

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

On the optimality of regularity in mixing Markovian decision rules for MDP control

Dinard van der Laan *

March 23, 2010

Abstract

In this paper we study Markov Decision Process (MDP) problems with the restriction that at decision epochs only a finite number of given Markovian decision rules may be applied. The elements of the finite set of allowed decision rules should be mixed to improve the performance. The set of allowed Markovian decision rules could for example consist of some easy-implementable decision rules, but also many open-loop control problems can be modelled as an MDP for which the applicable decision rules are restricted. For various subclasses of Markovian policies methods to maximize the performance are obtained, analyzed and illustrated with examples. Advantages and disadvantages of optimizing over particular subclasses of applicable policies are discussed and optimal performances are compared. One of the main results gives sufficient conditions for the existence of an optimal Markovian policy belonging to the subclass of applicable policies having a so-called regular structure.

Keywords: Markov Decision Process; Mixing Decision Rules; Optimization; Regular Sequences.

1 Introduction

Markov decision processes (MDP) are a well established tool for optimizing the control of stochastic systems. To model a complex system as MDP is applied in for example telecommunication, manufacturing systems and call centers. Classically solving the MDP results in an (optimal) policy which for every system state yields a corresponding (optimal) control action. To implement this policy at any decision event the current system state has to be known (or determined) before the corresponding control action is chosen. In practice such implementation may not be convenient. Moreover, for complex systems with large

*Tinbergen Institute, and Department of Econometrics and Operations Research, VU University, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands, Email: dalaan@feweb.vu.nl

(multi-dimensional) state space it is hard and practically impossible to find the optimal state-dependent policy.

In this paper we consider MDP for which decisions should be taken at an infinite discrete set T of consecutive decision epochs. For such MDP a general (possibly non-stationary) Markovian policy specifies for each decision epoch $t \in T$ a decision rule to be applied at t where a decision rule can be represented by a mapping from the state space to a corresponding action space. For purposes in this paper all of these spaces may assumed to be finite. Still in general it soon becomes intractable to determine the optimal policy if the state and/or action space(s) get larger. Moreover, optimal decision rules may have a complicated structure and be hardly implementable in practice. Therefore a basic idea in this paper is to optimize over a (much) smaller set of Markovian policies by rigorously restricting the set of allowed decision rules and thus the corresponding mappings from state space to action space are also restricted and for example have a specific structure. Let $\mathcal{D}$ be the set of decision rules (mappings) to which is restricted. To obtain useful results for implementation in practice $\mathcal{D}$ should be chosen such that some specific properties are satisfied. First $\mathcal{D}$ should only consist of easy implementable decision rules such that the implementation of any corresponding restricted Markovian policy is not a problem. A second objective is the set $\mathcal{D}$ being small such that optimization over Markovian policies with decision rules restricted to $\mathcal{D}$ becomes tractable. Additionally we would like that the performance obtained by optimization over Markovian policies with decision rules restricted to $\mathcal{D}$ is competitive to performances of other implementable policies which can be quickly found by for example applying some heuristic.

A reasonable and easy implementable decision rule which could be such an element of $\mathcal{D}$ could for example for an MDP associated with some routing problem with parallel servers be the rule that routes any arriving job at the moment of arrival to the server with the shortest expected remaining workload. This is a rule which for many such routing problems gives a reasonable performance, but possibly if it is combined with a few other decision rules in $\mathcal{D}$ a better performance could be obtained while optimization remains tractable.
If not performance but implementation is the issue for some MDP then arguably the most easy implementable decision rules are the rules which are given by a contant mapping. For such decision rules the chosen action at some decision epoch will not depend on the state of the stochastic process. This kind of mechanism to choose an action is also called state-independent or in a more general setting open-loop control. In fact many problems with open-loop control or partially observable MDP can be been as special cases of optimization over Markovian policies with decision rules restricted to $\mathcal{D}$ in which $\mathcal{D}$ consists only of decision rules corresponding to for example constant mappings or more generally mappings which are constant on given subdomains of the state space.

In this paper the problem of MDP optimization over Markovian policies with decision rules restricted to some given finite set $\mathcal{D}$ will be referred to as $\mathcal{D}$ restricted MDP. Policies which are applicable to $\mathcal{D}$ restricted MDP will be referred to as $\mathcal{D}$-mixing policies. First optimization over the class of stationary randomized $\mathcal{D}$-mixing policies is investigated and after that

optimization over certain classes of (possibly non-stationary) deterministic $\mathcal{D}$-mixing policies. The advantages and disadvantages of optimizing over particular classes of $\mathcal{D}$-mixing policies are discussed. Then for any $\mathcal{D}$ restricted MDP an associated MDP is defined for which the decision rules are not restricted and the action space is very simple, but the state space is continuous and thus much larger. The associated MDP is defined such that optimization is equivalent to optimization of the $\mathcal{D}$ restricted MDP and (optimal) sample paths of the former yield (optimal) $\mathcal{D}$-mixing policies. Such associated MDP will be referred to as full observation MDP. Although for these full observation MDP there is no longer a restriction on the decision rules obtaining an optimal policy remains difficult in practice mainly due to the (large) continuous state space of such full observation MDP. However, for the full observation MDP some well-known structural results like for example existence of a stationary and deterministic Markovian policy should hold if some appropriate conditions are satisfied. We will investigate whether such structural results on optimal policies and conditions to be satisfied for the full observation MDP have corresponding implications for the equivalent $\mathcal{D}$ restricted MDP. Indeed it turns out that for $\mathcal{D}$ restricted MDP also structural results on optimal policies can be formulated if certain conditions are satisfied, but that both structure and conditions are usually more complex than for MDP without restriction on the applicable decision rules.

One of the main results in this paper will be the existence of optimal deterministic $\mathcal{D}$-mixing policies having a so-called regular structure if some conditions are satisfied. The existence of optimal policies with a regular structure for some open-loop control problems is investigated before and [1] gives an overview on this. The obtained results were for queueing networks assumed to have particular topological properties. Moreover, in [1] the main condition for this optimality of a regular policy is multimodularity of the performance function. For many problems this condition of multimodularity is hard to check. In the present paper completely different conditions are obtained which are sufficient for the existence of a regular policy which is optimal. These conditions are formulated in terms of $\mathcal{D}$ restricted MDP and the associated full observation MDP and can also be checked for other type of problems then open-loop control in particular queueing networks.

This paper is organized as follows. Section 2 starts with some basic notation and MDP concepts and a performance measure is defined. Then the concepts of $\mathcal{D}$ restricted MDP and $\mathcal{D}$-mixing policies are introduced and explained in detail. The link between these concepts and open-loop control mechanisms is also explained. In Section 3 Bernoulli policies are introduced as a class of $\mathcal{D}$-mixing policies having the properties to be randomized and stationary. A method is given to compute the performance of any Bernoulli policy to be applied to a $\mathcal{D}$ restricted MDP. Also the problem of optimizing over the Bernoulli policies is considered and we give a method to do this. This method to optimize is illustrated with an example on a particular $\mathcal{D}$ restricted MDP. In Section 4 deterministic $\mathcal{D}$-mixing policies are introduced. Comparing with Bernoulli policies the advantages and disadvantages of applying and optimizing over deterministic $\mathcal{D}$-mixing policies are discussed. The problem of computing the performance of a given deterministic $\mathcal{D}$-mixing is considered and for so-called periodic policies a method is given and illustrated with an example.
In Section 5 for any given $\mathcal{D}$ restricted MDP the associated (full observation) MDP is defined such that there is equivalence with the $\mathcal{D}$ restricted MDP. This equivalence gives some

useful results on optimal policies and associated sample paths. Then some conditions for $\mathcal{D}$ restricted MDP are given which are shown to be sufficient for the existence of optimal stationary deterministic Markovian policies for the associated full observation MDP. Moreover, it is shown that these conditions are sufficient for performances of deterministic $\mathcal{D}$-mixing policies to be independent of the initial state distribution from which additional results are deduced. In Section 6 some special subclasses of deterministic $\mathcal{D}$-mixing policies are introduced. For the introduced subclasses algorithms are given to optimize the performance over such a subclass. The complexity of these algorithms is considered. A subclass of deterministic $\mathcal{D}$-mixing policies which is considered in particular are the so-called policies with regular structure for which the corresponding sequence of decision rules is a so-called regular sequence. The most sophisticated algorithm introduced in this section to optimize the performance over this subclass is illustrated in an example. In this section also some partial result is obtained on the optimal performance over some other subclass of policies compared to the optimal performance over all $\mathcal{D}$-mixing policies. To obtain this result it is again useful to consider the associated full observation MDP.
Finally in Section 7 it is proved that for $\mathcal{D}$ restricted MDP with $\mathcal{D}$ consisting of (only) two different decision rules some generally applicable conditions are sufficient for the existence of an optimal $\mathcal{D}$-mixing policies within the subclass of deterministic $\mathcal{D}$-mixing policies with regular structure. The associated full observation MDP is considered to formulate the main condition for this result. It is shown that for the full observation MDP the existence of an optimal stationary deterministic Markovian policy having some type of threshold structure is together with some easy checkable minor conditions sufficient to obtain the result on optimality within the sublass of policies with regular structure. The application of this result is illustrated with an example. Some concluding remarks are made about possible generalizations of the main result and possible connections with comparable MDP or optimal control problems.

2 Mixing of MDP decision rules

To describe the problem we investigate and corresponding results we first recall some basic notation and definitions for MDP's. As far as possible the MDP notations and definitions from [21] are followed. We also introduce some additional notations, definitions for specifically needed for this paper, meanwhile explaining the concept of mixing decision rules and corresponding optimization problems.

We recall some MDP notation following [21], meanwhile mentioning for example specific assumptions, restrictions, etc. for this paper. Let T be the set of decision epochs and S be the state space. Unless T is explicitly given we assume in this paper T to be discrete and we assume an infinite horizon MDP, say $T = \{1, 2, \ldots\}$. Moreover, S is assumed to be finite. Most results in this paper extend to more general state space, but we focus on the basis ideas and more generality could distract (for example existence issues). Let A_s be the action space for state $s \in S$ and $A = \cup_{s \in S} A_s$ which is also assumed to be finite. Special fo-

cus will be for the case that $A = A_s$ for every $s \in S$, i.e. a common action space for all states.

If in state s at decision epoch t action $a \in A_s$ is chosen, then an (expected) immediate reward $r_t(s,a)$ is received and the system state at the next decision epoch is determined by the probability distribution $p_t(\cdot|s,a)$. In this paper we assume stationary stationary rewards and transition probabilities. In other words these immediate rewards and probability distributions will not depend on t and therefore we may omit the subscript t in notation. Moreover, we assume bounded rewards. Thus $|r(s,a)| \leq M < \infty$ for all $a \in A_s$ and $s \in S$. If the immediate reward also depends on the transition to the next state and $r_(s,a,j)$ denotes the reward received if by choosing action a a transition is made from state s to state j then for all s, $a \in A_s$ the expected immediate rewards may be evaluated by computing

$$r(s,a) = \sum_{j \in S} r(s,a,j) p(j|s,a).$$

The collection of objects $\{T, S, A_s, p_t(\cdot|s,a), r_t(s,a)\}$ is referred to as Markov decision process (MDP). Together with an optimality criterion it is also called a Markov decision problem. Slightly abusing notation in this paper MDP may be used for both.

2.1 Decision rules and policies

A decision rule prescribes at a specific decision epoch for each state $s \in S$ a procedure for selecting an action $a \in A_s$. A decision rule is Markovian (memoryless) if only the current system state at the decision epoch may be used to determine which action is chosen and a decision rule is deterministic if it chooses an action with certainty. In this paper the main focus is on Markovian deterministic (MD) decision rules, which are generally the most easiest to implement. However, also Markovian randomized (MR) decision rules will be considered. An MD decision rule for a particular $t \in T$ is equivalent to a mapping $d_t : S \to A_s$ specifying the action to be chosen in state s at decision epoch t. On the other hand MR decision rules map the set of states into the set of probability distributions on the action space, that is $d_t : S \to \mathcal{P}(A)$.

A policy (strategy) specifies for all decision epochs t which decision rule d_t is used. Thus in this paper considering infinite horizon MDP's a policy π is determined by an infinite sequence $\pi = (d_1, d_2, \ldots)$. If $d_t \in MD$ ($d_t \in MR$) for all $t \in T$ then π is said to be Markovian deterministic ($\pi \in \Pi^{MD}$) respectively Markovian randomized ($\pi \in \Pi^{MR}$). Policy π is called stationary if $d_t = d$ for all $t \in T$. Then Π^{SD} denotes the class of policies which are both stationary and Markovian deterministic which is a subset of Π^{SR}, the class of policies which are both stationary and Markovian randomized.

2.2 Induced Markov chains and optimization over mixing policies

Let $\Omega = \{S \times A\}^\infty$ be the sample space for the stochastic process generated by the MDP applying some policy π. An element $\omega \in \Omega$ is an alternating sequence $\omega = (s_1, a_1, s_2, a_2, \ldots)$ of states and actions which is referred to as a sample path. For all $t \in T$ random variables X_t and Y_t are defined by $X_t(\omega) = s_t$ and $Y_t(\omega) = a_t$. Thus X_t denotes the state at t and Y_t the action chosen at decision epoch t. Since we restrict to Markovian policies π it follows that the induced stochastic process $\{X_t; t \in T\}$ is a discrete time Markov chain. This Markov chain is stationary if π is also stationary. The bivariate stochastic process $\{X_t, r_t(X_t, Y_t)\}$ is referred to as Markov reward process. In this paper the optimality criterion is the lim inf average reward criterion. In other words for initial probability distribution x on state space S the performance of policy π is given by

$$g^\pi(x) := \liminf_{N\to\infty} \frac{1}{N} \mathbb{E}_x^\pi \left\{ \sum_{t=1}^{N} r(X_t, Y_t) \right\}. \tag{1}$$

In this paper we would like to restrict to problems for which $g^\pi(x)$ does not depend on the initial distribution x and then the performance of policy π may simply be denoted by g^π. For now we assume this property holds and later we give some explicit and sufficient conditions for this.

The main issue in this paper is optimization over policies for which all the decision rules d_t, $t = 1, 2, \ldots$ are restricted to be elements of some finite set of particular MD decision rules. Thus the problem is to maximize for any initial state distribution x the performance $g^\pi(x)$ over policies $\pi = (d_1, d_2, \ldots)$ restricted to $d_t \in \mathcal{D}$ for every $t \in T$ where $\mathcal{D}$ is a given finite set of MD decision rules.

In practical applications the given set $\mathcal{D}$ of allowed decision rules typically consist of easy implementable MD decision rules determined by some straightforward heuristic. For example in a routing problem such heuristic rules could be "route arriving jobs to the shortest queue" or "route arriving jobs to the queue being served by the fastest server". In general such rules are suboptimal with respect to performance optimization.

In the sequel we assume that the set of "allowed" decision rules $\mathcal{D}$ consists of two MD decision rules d^1 and d^2. Then the only two (Markovian) stationary deterministic policies for which decision rules for all $t \in T$ are restricted to $\mathcal{D}$ are the policies $\pi^1 := (d^1, d^1, \ldots) \in \Pi^{MD}$ and $\pi^2 := (d^2, d^2, \ldots) \in \Pi^{MD}$. Since π_1 and π_2 are stationary policies they induce stationary discrete time Markov chains on S with corresponding transition matrices P and Q respectively. We assume that both Markov chains are unichain and aperiodic. In other words both Markov chains have exactly one recurrent class which is aperiodic and let p, q be the corresponding unique stationary distributions satisfying $p^T = p^T P$, $\sum_{s\in S} p_s = 1$ respectively $q^T = q^T Q$, $\sum_{s\in S} q_s = 1$, where p^T and q^T are the row vectors representing the stationary distributions p and q respectively. The finiteness of S guarantees the existence of p and q and the performances g^{π_1}, g^{π_2} of both policies may be directly computed from p and q

respectively. From the existence of such unique stationary distributions p and q it follows that the performances g^{π_1} and g^{π_2} of the two stationary policies are independent of the initial state distribution. Indeed for all initial state distributions on S the performance of policy π_1 is given by $g^{\pi_1} = \sum_{s \in S} p(s) r(s, d^1(s))$. Similarly $g^{\pi_2} = \sum_{s \in S} q(s) r(s, d^2(s))$ gives the performance of policy π_2.

We may generalize these formulas to compute the performance of any (randomized) stationary policy $\pi = (d, d, \ldots) \in \Pi^{MR}$ where d is some randomized decison rule inducing a stationary unichain aperiodic Markov chain. Indeed let p be the unique stationary distribution on state space S of the induced Markov chain and $r(d)_s := \sum_{a \in A_s} r(s,a) Pr(a|s,d)$ the expected immediate reward in state $s \in S$ given that MR decision rule d is applied. Then the expected performance of π is given by

$$g^{\pi} = \sum_{s \in S} p_s r(\pi)_s = p \cdot r(d), \text{ the inner product of } p \text{ and } r(d). \tag{2}$$

However, for non-stationary policies $\pi = (d_1, d_2, \ldots)$ with $d_t \in \mathcal{D}$ for $t = 1, 2, \ldots$ the performance may depend on the initial distribution even if all transition matrices corresponding to decision rules in $\mathcal{D}$ are unichain and aperiodic. Example 11 will show this. In Section 5 some more (additionally to unichain and aperiodic) conditions on the transition matrices will be given such that also for non-stationary Markovian policies π the performance will not depend on the initial state distribution.

Let $\pi^1, \pi^2 \in \Pi^{MD}$ be both stationary deterministic policies for $\mathcal{D} = \{d^1, d^2\}$ as above. Both π^1 and π^2 are not optimal if d^1 respectively d^2 are suboptimal decision rules. In this case we will see that despite having the restriction $d_t \in \{d^1, d^2\}$ for all $t \in T$ for Markovian policies $\pi = (d_1, d_2, \ldots)$ the performance could be improved if π is not required to be both stationary and deterministic. Within this set of allowed policies the objective is to maximize the performance g^{π} (assuming for now it does not depend on the initial state distribution) resulting in a performance that is strictly larger than $\max(g^{\pi_1}, g^{\pi_2})$. For such a policy π the MD decision rules d^1 and d^2 have to be *mixed* in some way and therefore we call the considered policies $\mathcal{D}$-mixing policies.

To explain the basic concepts and optimization tools we investigate the case that $\mathcal{D}$ consists of two (suboptimal) MD decision rules, but we note that the mixing of decision rules can be applied similarly if $\mathcal{D}$ consists of more than two MD decision rules. Also mixing the more general MR decision rules instead of MD decision rules is theoretically not a problem, but in practice and the examples we discuss in this paper the set $\mathcal{D}$ consists of only MD decision rules. In the following sections we discuss in detail the concept of mixing decision rules and several approaches to improve the performance in this way. Structural results for optimization over $\mathcal{D}$-mixing policies will be derived.

2.3 Open-loop control and corresponding mixing policies

For many applications performance optimization yield an MDP for which it is desirable to use an open-loop control mechanism. In this case the choice of an action should not depend on the (current) system state. For example if at decision epochs observing the current system state is relatively expensive, time-consuming or not possible at all then open-loop (state-independent) control should be considered. To apply open-loop control we assume that there is a common action space A for all states. The most simple case is $A = \{a, b\}$, i.e. in every state the same two actions a and b are available. For example in a queueing problem with admission control with decision epochs corresponding to arrivals of jobs action a could be to accept the new arriving job and action b to decline it. If $A = \{a, b\}$ then the only two decision rules which obey the rules of open-loop control are d^1 which chooses action a in every state $s \in S$ and d^2 which chooses action b in every state $s \in S$. Analogously to the policy mixing over two MD decision rules described in the previous subsection d^1 induces a stationary Markov chain with some corresponding transition matrix P and d^2 induces a stationary Markov chain with some corresponding transition matrix Q. Moreover, any (Markovian) open-loop control policy π is of the form $\pi = (d_1, d_2, \ldots)$ with $d_t \in \{d^1, d^2\}$ for every $t \in T$ and it follows that optimizing the performance over all open-loop control policies with two available actions in every state can be considered as a special case of optimization over $\{d^1, d^2\}$-mixing policies as described in the previous subsection. Similarly optimizing open-loop control with any finite common action space A corresponds to optimization over $\mathcal{D}$-mixing policies where $\mathcal{D}$ has the same cardinality as the action space A.

3 Bernoulli policies

Given some MDP and $\mathcal{D} = \{d^1, d^2\}$ a set of two allowed MD decision rules for controlling the MDP as in the previous section. Consider the following randomized algorithm to generate $\mathcal{D}$-mixing policies $\pi = (d_1, d_2, \ldots)$. Let $\theta \in [0, 1]$ be given. For $t = 1, 2, \ldots$ generate independent random numbers u_t uniformly distributed on $[0, 1]$ and put $d_t = \begin{cases} d^1 & \text{if } u_t \in [0, \theta] \\ d^2 & \text{if } u_t \in (\theta, 1] \end{cases}$. In other words for every decision epoch $t \in T$ an independent θ-coin is flipped which outcome determines the decision rule which is applied at t. For all $t \in T$ with probability θ the first decision rule is applied and with probability $1 - \theta$ the second decision rule. Policies generated by this randomized algorithm are called Bernoulli policies of rate θ. Note that above Bernoulli algorithm may easily be generalized for the case that $\mathcal{D}$ consists of more than two decision rules, but this generalization is not explored in this paper.

The randomization of the policy in the Bernoulli algorithm makes actual implementation of such policies in practice somewhat awkward, but a nice property is that the performance of Bernoulli policies is relatively easy to compute or approximate. This makes it tractable to optimize the performance over all Bernoulli policies and in particular analytic methods are available for optimizing the Bernoulli parameter θ. The following property of the Bernoulli policy is important to analyze and compute or approximate performances.

Lemma 1 *Assume an MDP with finite state space S where decisions rules d^1 and d^2 induce stationary and aperiodic unichain Markov chains with corresponding transition matrices P respectively Q. Then any Bernoulli policy mixing d^1 and d^2 with rate $\theta \in [0,1]$ induces a stationary aperiodic unichain Markov chain on S with transition matrix*

$$B_\theta = \theta P + (1-\theta)Q \tag{3}$$

which has an unique stationary distribution b_θ satisfying $b_\theta^T B_\theta = b_\theta^T$ and $\sum_{s\in S} b_\theta(s) = 1$, where b_θ^T is the row vector representing b_θ.

In other words the Bernoulli policy mixing two decision rules induces a stationary Markov chain with unique stationary distribution b_θ depending on the Bernoulli parameter θ. From this it follows that given the MDP and decision rules d^1 and d^2 the expected performance of the Bernoulli policy is a function $g(\theta)$ of the Bernoulli parameter θ and by (2) we have

$$g(\theta) = \sum_{s\in S}(b_\theta)_s[\theta r(s,d^1(s)) + (1-\theta)r(s,d^2(s))] = \theta(b_\theta \cdot r(d^1)) + (1-\theta)(b_\theta \cdot r(d^2)). \tag{4}$$

3.1 Optimizing over Bernoulli policies

Optimizing the expected performance $g(\theta)$ of Bernoulli policies over $\theta \in [0,1]$ is relatively easy if $g(\theta)$ is a smooth function of $\theta \in [0,1]$. Indeed the function $g(\theta)$ is usually smooth on the interval $[0,1]$ and for example from [11] follow sufficient conditions for the smoothness of $g(\theta)$. To give a sufficient condition for the smoothness we first combine some notations, definitions and results from [11] and [21]. Let P be a transition matrix of an irreducible finite state Markov chain. The stationary (limiting) matrix P^* of P is defined by the Cesaro limit

$$P^* = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} P^n.$$

P^* has all its rows equal to p^T, the row-vector of the stationary distribution p of P. The deviation matrix D_P of P is defined by

$$D_P = (I - P + P^*)^{-1}(I - P^*)$$

which equals the Cesaro limit for $N \to \infty$ of $\sum_{n=0}^{N-1}(P^n - P^*)$. These are all square matrices in $\mathbb{R}^{|S|\times|S|}$ and the so-called maximum absolute row sum norm $||\quad||_\infty$ is defined by

$$||B||_\infty = \max_{i\in S} \sum_{j\in S} |B(i,j)| \text{ for } B \in \mathbb{R}^{|S|\times|S|}$$

Let transition matrices P and Q be as in Lemma 1. Then with above notations and definitions a sufficient condition (C) folowing from [11] for the smoothness of the performance function $g(\theta)$ is the following.

Condition 2 *There exists some integer* $n \in \mathbb{N}$ *such that*

$$||((P-Q)D_Q)^n||_\infty < 1.$$

Combining Lemma 3.1 and Lemma 3.3 from [11] the following result follows.

Theorem 3 *Let* $P, Q \in R^{|S|\times|S|}$ *be as in Lemma 1. For all* $\theta \in [0,1]$ *let* $B_\theta = \theta P + (1-\theta)Q$ *and* B_θ^* *the stationary matrix of* B_θ*. Moreover, let* $I \in R^{|S|\times|S|}$ *be the identity matrix and put* $A := (P-Q)D_Q$*. Suppose condition 2 is satisfied. Then for any* $\theta \in [0,1]$ *the Neumann series expansion* $\lim_{k\to\infty}\sum_{n=0}^{k}((B_\theta - Q)D_Q)^n = \lim_{k\to\infty}\sum_{n=0}^{k}(\theta A)^n$ *is convergent and*

$$B_\theta^* = Q^*(I-\theta A)^{-1} = Q^* \lim_{k\to\infty}\sum_{n=0}^{k}(\theta A)^n = Q^* \lim_{k\to\infty}\sum_{n=0}^{k}((B_\theta - Q)D_Q)^n. \tag{5}$$

From (5) it follows that any component of the stationary distribution b_θ of B_θ which is an (arbitrary) row of B_θ^* is a smooth function of θ for $\theta \in [0,1]$. Combining this with equation (4) gives the following result.

Corollary 4 *If Condition 2 is satisfied then the performance function* $g(\theta)$ *is smooth on the interval* $[0,1]$*. Moreover, there exists some* $\theta^* \in [0,1]$ *maximizing the performance* $g(\theta)$ *of Bernoulli policies and any optimal* θ^* *satisfies (at least) one of the following conditions:*

- $\theta^* = 0$
- $\theta^* = 1$
- $g'(\theta^*) = 0$

Remark 5 We have given a sufficient Condition 2 only for the $||\quad||_\infty$ matrix norm, but it may be generalized to so-called v-norms as in [11]. Moreover, for applications in this paper the role of matrices P and Q are symmetric and may be interchanged to get an alternative sufficient condition which has the same implications as Theorem 3 except for the interchange of the roles of P and Q and Condition 6 is therefore also sufficient for the properties of optimal θ^* as in Corollary 4. This alternative sufficient Condition 6 is given below and in fact in this condition the role of P and Q is exactly the same as in [11].

Condition 6 *There exists some integer* $n \in \mathbb{N}$ *such that*

$$||((Q-P)D_P)^n||_\infty < 1.$$

3.2 Example

In this subsection we treat an example with a state space consisting of only two states, but for which the application and optimization of Bernoulli mixing policies is non-trivial and previous results as Corollary 4 are well illustrated. In fact in the sequel of this paper the same MDP example will be used to illustrate also optimization over $\mathcal{D}$-mixing policies of another type than the Bernoulli policies discussed in this section and results will be compared.

Example 7 In this example a machine is operated which can be in two states, state space $S = \{1, 2\}$, where state 1 is referred to as the bad state and state 2 as the good state. At every decision epoch t, $t = 1, 2, \ldots$ the operator has to decide whether the machine goes in working or repair mode for one time-unit until the next decision epoch. Thus there is a common action space $\mathcal{A} = \{1, 2\}$ for both states where action 1 refers to working mode and action 2 refers to repair mode. If action 1 is chosen then there is a probability of 0.2 that the machine will be in bad state at the next decision epoch if the machine is currently in good state. Moreover, for action 1 the machine will allways be in bad state at the next decision epoch if the current state is bad. For action 2 there is a probability of 0.3 that the machine will be in good state at the next decision epoch if the machine is currently in bad state. Moreover, for action 2 the machine will allways be in good state at the next decision epoch if the current state is good. The only case in which a positive immediate reward of 1 is obtained if action 1 is chosen and the machine is currently in good state, in every other case we assume that the immediate reward is 0. For this MDP with average reward criterion finding the optimal policy is very easy to solve. Of course if the machine is in good state then action 1 will be optimal and if the machine is in bad state then action 2 is optimal. However, we consider optimization over $\mathcal{D} = \{d^1, d^2\}$-mixing policies where decision d^1 is choosing action 1 (work) for both states and d^2 is choosing action 2 (repair) for both states. Note that d^1 and d^2 are deterministic open-loop decision rules and $\{d^1, d^2\}$-mixing policies could be considered for example if observing the (current) state of the machine has some cost. Optimizing over $\{d^1, d^2\}$-mixing policies is not trivial for this problem and for now we consider optimization over all Bernoulli $\{d^1, d^2\}$-mixing policies.

For $\{d^1, d^2\}$-mixing policies the problem of maximizing the long-run average reward is completely specified by the transition matrix P and expected immediate reward vector $r(d^1)$ induced by decision rule d^1, respectively the transition matrix Q and expected immediate reward vector $r(d^2)$ induced by decision rule d^2. From the model description above it follows that for this example we have that

$$P = \begin{pmatrix} 1 & 0 \\ 0.2 & 0.8 \end{pmatrix}, \quad r(d^1) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad Q = \begin{pmatrix} 0.7 & 0.3 \\ 0 & 1 \end{pmatrix}, \quad r(d^2) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{6}$$

By Lemma 1 it follows that applying a Bernoulli policy of rate $\theta \in [0, 1]$ induces transition

matrix

$$B_\theta = \theta P + (1-\theta)Q = \begin{pmatrix} 0.7+0.3\theta & 0.3-0.3\theta \\ 0.2\theta & 1-0.2\theta \end{pmatrix} \tag{7}$$

it is easily checked that Condition 2 is satisfied for this example. Indeed

$$D_Q = \frac{1}{3}\begin{pmatrix} 10 & -10 \\ 0 & 0 \end{pmatrix}, \quad A := (P-Q)D_Q = \begin{pmatrix} 1 & -1 \\ \frac{2}{3} & -\frac{2}{3} \end{pmatrix} \text{ and thus}$$

$$A^2 = \begin{pmatrix} \frac{1}{3} & -\frac{1}{3} \\ \frac{2}{9} & -\frac{2}{9} \end{pmatrix} \text{ and } ||A^2||_\infty = \frac{2}{3} < 1.$$

Moreover, it is easily obtained that $b_\theta^T = (\frac{2\theta}{3-\theta}, \frac{3-3\theta}{3-\theta})$ is the row vector corresponding to the stationary distribution of B_θ for any $\theta \in [0,1]$. Using this we may obtain the performance function $g(\theta)$ by substituting $b(\theta)$, $r(d^1)$ and $r(d^2)$ in (2) and it follows that

$$g(\theta) = \theta(\frac{2\theta}{3-\theta}, \frac{3-3\theta}{3-\theta})\begin{pmatrix} 0 \\ 1 \end{pmatrix} + (1-\theta)(\frac{2\theta}{3-\theta}, \frac{3-3\theta}{3-\theta})\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \frac{3\theta - 3\theta^2}{3-\theta}.$$

Obviously $g(0) = g(1) = 0$ and $g(\theta) > 0$ for $\theta \in (0,1)$. Thus from Corollary 4 it follows that $g'(\theta^*) = 0$ for any optimal $\theta^* \in [0,1]$. Since $g'(\theta) = \frac{3(\theta^2-6\theta+3)}{(\theta-3)^2}$ it follows that $\theta^* = 3-\sqrt{6} \sim 0.551$ is the unique value for $\theta \in [0,1]$ that maximizes the performance of the corresponding Bernoulli policy of rate θ. The performance of this optimal Bernoulli policy equals $g(3-\sqrt{6}) = 15-6\sqrt{6} \sim 0.303$.

Remark 8 In case of complex systems for which the MDP has a very large state space it is not tractable to obtain exact expressions for the stationary distribution b_θ and performance function $g(\theta)$ as in the example above. However, also for very large state space an approximation of the stationary distribution b_θ may be obtained by methods like Markov chain Monte Carlo and then the expected performance $g(\theta)$ of the Bernoulli policy could also be approximated by plugging in the approximation of b_θ in (4). Thus the optimal θ^* maximizing $g(\theta)$ may also be approximated in such cases. Moreover, if $g(\theta)$ is differentiable then gradient estimation by measure valued differentiation could be applied to approximate some (optimal) value $\theta^* \in [0,1]$ for which $g'(\theta^*) = 0$. In [5] this simulation technique is applied to a call center operation problem with two types of jobs having different service requirements where in various ways two reasonable applicable decision rules are obtained which are mixed to improve the system performance. The technique is relatively fast to approximate an optimal value for θ. For more theoretical results and background on this see for example [10], [12] and [13].

4 Deterministic mixing policies

In the previous section we have investigated the optimization of Bernoulli policies which rely on a randomized mechanism to generate $\mathcal{D}$-mixing policies. However, in the present section

we investigate the optimization of deterministic $\mathcal{D}$-mixing policies which are represented as an infinite deterministic sequence describing for every decision epoch which (Markovian deterministic) decision rule in $\mathcal{D}$ is applied. For example if $\mathcal{D} = \{d^1, d^2\}$ and we let symbol 1 correspond to decision rule d^1 and symbol 0 to decision rule d^2 then we have a one-to-one correspondence between deterministic $\mathcal{D}$-mixing policies and one-sided infinite sequences $U = (u_1, u_2, \ldots)$ of zeros and ones. Therefore in the sequel an infinite sequence $(u_1, u_2, \ldots)$ is identified with a deterministic $\mathcal{D}$-mixing policy where u_t determines the decision rule which is applied at decision epoch t for $t = 1, 2, \ldots$. Thus if $\mathcal{D} = \{d^1, d^2\}$ then optimizing the performance over all deterministic $\mathcal{D}$-mixing policies corresponds to optimization over the set $\{0, 1\}^{\mathbb{N}}$ of all one-sided infinite sequences of zeros and ones. More generally if $\mathcal{D} = \{d^1, d^2, \ldots, d^n\}$ then it follows analogously that optimizing the performance over all deterministic $\mathcal{D}$-mixing policies corresponds to optimization over a corresponding set W where $W = \{\mathcal{A}\}^{\mathbb{N}}$ are all one-sided infinite words over alphabet $\mathcal{A}$. A word is by definition a sequence of symbols from a finite alphabet and if $\mathcal{D} = \{d^1, d^2, \ldots, d^n\}$ then the corresponding alphabet $\mathcal{A}$ consists of n (different) symbols.

4.1 Positive and negative aspects

One of the positive aspects of applying a deterministic $\mathcal{D}$-mixing policy which as above is represented as infinite decision sequence $U = (u_1, u_2, \ldots)$ is that the implementation is more straightforward than for Bernoulli policies. Indeed at decision epoch t only the (easy implementable) MD decision rule determined by u_t has to be implemented and it is not necessary to "flip a coin" (randomization) at every decision epoch as is the case for Bernoulli policies. Thus the "additional randomness" in the system evolution created by Bernoulli policies which may be a problem for operators in real-life applications is nonexistent for deterministic $\mathcal{D}$-mixing policies.
Moreover, arguably the most important advantage of deterministic mixing policies compared to Bernoulli policies is that in general good (not necessarily optimal) deterministic mixing policies easily outperform the best (optimized) Bernoulli policy. Note that implementing any Bernoulli policy creates a (random) infinite sequence of decision rules. Such an sequence is an element of W, while deterministic mixing policies may optimize over all elements of W. Thus an optimal deterministic mixing policy performs generally better than an optimal Bernoulli policy. Thus there are several advantages of deterministic mixing policies compared with Bernoulli policies.
There are also some disadvantages which we discuss now. One of the nice properties of applying a Bernoulli mixing policy is that it induces a stationary Markov chain on the state space S. In contrast deterministic mixing policies given by some infinite decision sequence $U = (u_1, u_2, \ldots)$ do not induce a stationary Markov chain except for degenerate policies for which $u_t = u_1$ for every decision epoch t. Therefore it is also not possible to obtain the performance of deterministic mixing policies by computing an unique stationary distribution and applying (4) as for Bernoulli policies. The fact that computing the performance is harder than for Bernoulli policies is one of the reasons that optimizing over deterministic mixing policies is also much harder than for Bernoulli policies. For periodic deterministic

mixing policies there exists an algorithm to compute the performance, but computation times will increase with the period. A deterministic mixing policy is periodic with period k if for the corresponding decision sequence $U = (u_1, u_2, \ldots)$ it holds that $u_t = u_{t+k}$ for $t = 1, 2, \ldots$. Below we give an example illustrating the computation of the performance of periodic deterministic mixing policies.

Example 9 Consider again Example 7 from the previous section which characteristics were summarized by (6). Instead of the performance of Bernoulli policies we now compute the performance of the deterministic mixing policy π with corresponding decision sequence $U = (1, 0, 1, 0, \ldots) = (1, 0)^\infty$ which obviously is periodic with period 2. For $t = 1, 2, \ldots$ let $X_t \in \{1, 2\}$ be the state at decision epoch t when policy π is applied. Then $\{X_t, t = 1, 2, \ldots\}$ is a Markov chain which is not stationary. However, $\{X_t, t = 1, 3, 5, \ldots\}$ is a stationary Markov chain with transition matrix

$$A_1 = PQ = \begin{pmatrix} 0.7 & 0.3 \\ 0.14 & 0.86 \end{pmatrix}.$$

It is easily verified that this Markov chain has unique stationary distribution $b_1^T = (\frac{7}{22}, \frac{15}{22})$. Analogously $\{X_t, t = 2, 4, 6, \ldots\}$ is a stationary Markov chain with transition matrix

$$A_2 = QP = \begin{pmatrix} 0.76 & 0.24 \\ 0.20 & 0.80 \end{pmatrix}$$

and unique stationary distribution $b_2^T = (\frac{5}{11}, \frac{6}{11})$. It follows that for $t = 1, 3, \ldots$ the long-run average reward is given by the inner product $b_1 \cdot r(d^1) = \frac{15}{22}$ and for $t = 2, 4, \ldots$ the long-run average reward is given by the inner product $b_2 \cdot r(d^2) = 0$. This implies that for the performance $g(\pi)$ we have that $g(\pi) = \frac{1}{2}(\frac{15}{22} + 0) = \frac{15}{44} \sim 0.341$.

In Example 9 the performance of a deterministic mixing policy with period 2 is computed. The following theorem generalizes the performance formula for any periodic deterministic mixing policy.

Theorem 10 *Let π be a deterministic $\mathcal{D}$-mixing policy with corresponding decision sequence $U = (u_1, u_2, \ldots)$ which is periodic with period k. Let X_t be the state at decision epoch t when policy π is applied and $d_t \in \mathcal{D}$ the decision rule corresponding to u_t to be applied at decision epoch t. For $m = 1, 2, \ldots, k$ assume that the stationary Markov chain $\{X_t, t = m, m + k, m + 2k, \ldots\}$ has unique stationary distribution b_m. Then for the long-run average reward $g(\pi)$ we have that*

$$g(\pi) = \frac{1}{k} \sum_{m=1}^{k} b_m \cdot r(d^m). \tag{8}$$

Note that according to (8) the performance $g(\pi)$ of such a periodic policy π does not depend on the initial state distribution. However, in Theorem 10 the assumption is made that for

all m the stationary Markov chain $\{X_t, t = m, m+k, m+2k, \ldots\}$ has an unique stationary distribution. It is important to realize that this is a necessary condition and that for this condition to hold it is not sufficient that all the transition matrices associated to the decision rules in $\mathcal{D}$ are unichain and aperiodic. Indeed we have the following counterexample.

Example 11 Suppose again $\mathcal{D} = \{d^1, d^2\}$ as in Example 9 and the same periodic determistic $\mathcal{D}$-mixing policy π as in Example 9 corresponding to decision sequence $U = (1, 0, 1, 0, \ldots) = (1, 0)^\infty$ is applied. Let

$$P = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \text{ and } Q = \begin{pmatrix} 0 & 0 & 1 \\ 0.5 & 0 & 0.5 \\ 0 & 1 & 0 \end{pmatrix}$$

be the transition matrices corresponding to decision rules d^1 respectively d^2. It is easily seen that P and Q are unichain and aperiodic. However,

$$A_1 := PQ = \begin{pmatrix} 0.25 & 0.5 & 0.25 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

is obviously not unichain and thus the Markov chain $\{X_1, X_3, X_5, \ldots\}$ does not have an unique stationary distribution. Similarly

$$A_2 := QP = \begin{pmatrix} 1 & 0 & 0 \\ 0.5 & 0.25 & 0.25 \\ 0 & 0 & 1 \end{pmatrix}$$

is not unichain and thus also the Markov chain $\{X_2, X_4, X_6, \ldots\}$ does not have an unique stationary distribution. In fact in this example the performance of policy π in general depends on the initial state distribution and can not be computed by (8).

In Section 5 we will give some conditions on the transition matrices corresponding to the decision sequences in $\mathcal{D}$ which are sufficient for the independence of the performance of $\mathcal{D}$-mixing policies on the initial state distribution and under such conditions (8) is certainly valid.

We also note that (8) could be seen as a generalization of (2). Indeed the latter formula is then for the special case $k = 1$. Also it follows that if (8) is applied to compute the performance that the computational effort increases with the period k of the decision sequence U. In fact it is easily seen that for a fixed set $\mathcal{D}$ of allowed decision rules the computation time of computing the performance of a periodic deterministic $\mathcal{D}$-mixing policy by applying (8) increases linearly with the period k of the decision sequence.

Recall from Example 7 that for the performance of Bernoulli policies we obtained a closed formula in the parameter θ. More generally, for Bernoulli policies the computational effort to obtain the performance does hardly depend on which Bernoulli policy is applied since

all Bernoulli policies induce stationary Markov chains. Thus in comparison with Bernoulli policies computing the performance is harder for deterministic mixing policies since in general such policies do not induce stationary Markov chains. Moreover, for periodic policies the computational effort increases with the period. Even more problematic are deterministic mixing policies which are not periodic since in that case we do not have an algorithm to compute the exact performance (despite assuming it to independent of the initial state distribution) in finite time and we think it is in general only possible to approximate the performance of such a policy. Therefore the optimization over deterministic mixing policies is harder than optimization over all Bernoulli policies as we did in Example 7. Another issue is that deterministic $\mathcal{D}$-mixing policies may be optimized over the infinite discrete set W of all possible decision sequences $U = (u_1, u_2, \ldots)$ with $u_n \in \mathcal{D}$ for $n = 1, 2, \ldots$ which structure is more complicated than for Bernoulli policies for which the optimization is over a compact set. Recall for example from 7 that we could optimize over all Bernoulli policies by optimizing the parameter $\theta \in [0, 1]$.

Summarized a disadvantage of considering deterministic mixing policies is that the performance computation and optimization is more complicated than for Bernoulli policies. On the other hand the advantage of considering deterministic mixing policies is that many of them perform better than the optimal (and thus any) Bernoulli policy. For example the performance of the periodic policy considered in Example 9 is about 0.341 which improves the performance of the optimal $\mathcal{D}$ -Bernoulli mixing policy which is about 0.303 (see Example 7). We will see that the policy considered in Example 9 is not yet optimal within the class of deterministic mixing policies and thus the performance of 0.341 could be improved. However, optimization over the whole set W of possible deterministic decision sequences is intractable and therefore our approach in Section 6 will be to optimize over some specific subset(s) of W. The idea is to consider subsets of decision sequences with a specific structure.

5 The associated MDP

In this section we define an equivalent MDP associated to optimizing over $\mathcal{D}$-mixing policies for $\mathcal{D} = \{d^1, d^2\}$. The advantage of considering the equivalent MDP is that policies are no longer restricted to the class of $\mathcal{D}$-mixing policies. Therefore in contrast to the class of $\mathcal{D}$-mixing policies the existence of optimal Markovian policies which are both stationary and deterministic holds if certain conditions are satisfied. Then the existence of an optimal stationary deterministic Markovian policy for the equivalent MDP may be used to obtain structural properties of some optimal policy within the class of $\mathcal{D}$-mixing polcicies. In this way we will obtain results about optimality within certain subclasses of $\mathcal{D}$- mixing policies if some conditions are satisfied. Besides such benefits the equivalent MDP formulation also gives complications. For example in the equivalent MDP the state space is not finite anymore, but continuous. Practically this means that also for such associated MDP it is very hard to obtain an optimal solution. Thus it is hardly possible to obtain an optimal $\mathcal{D}$-mixing policy via this associated MDP approach, but structural results may be obtained.

Before we give the equivalent MDP to optimizing within the class of $\mathcal{D} = \{d^1, d^2\}$-mixing policies we recall some definitions and notations. Let S be the finite state space and $r(d^1)$ ($r(d^2)$) be the immediate reward vector for decision rule d^1 respectively d^2. Moreover, let P be the transition matrix associated to d^1 and Q be the transition matrix associated to d^2. Then an equivalent MDP with continuous state space is defined as follows:

- The state space X is the set of all probability distributions on S.
- The action space $\widetilde{A} := \{d^1, d^2\}$ for all $x \in X$.
- For all $x \in X$ the immediate rewards $r(x, d^1)$ and $r(x, d^2)$ are given by the inner products $r(x, d^1) := x \cdot r(d^1)$ respectively $r(x, d^2) := x \cdot r(d^2)$.
- For action d^1 state transitions are given by the state space mapping $x \to xP$ for all $x \in X$ where x is represented as $|S|$-dimensional row vector. Analogously for d^2 state transitions are given by the state space mapping $x \to xQ$ for all $x \in X$ being represented by a row vector.
- Let $\widetilde{\Omega} = \{X \times \widetilde{A}\}^\infty$ be the sample space for the stochastic process generated by the MDP when some admissible policy $\widetilde{\pi}$ is applied. A sample path $\widetilde{\omega} \in \widetilde{\Omega}$ is an alternating sequence $\widetilde{\omega} = (x_1, a_1, x_2, a_2, \ldots)$ of states and actions. For $t = 1, 2, \ldots$ let variables $\widetilde{X}_t$ and $\widetilde{Y}_t$ be defined by $\widetilde{X}_t(\omega) = x_t$ and $\widetilde{Y}_t(\omega) = a_t$. The optimality criterion is again the lim inf average reward criterion. In other words for initial state $\widetilde{X_1} = x \in X$ the performance of policy $\widetilde{\pi}$ is given by

$$g^{\widetilde{\pi}}(x) := \liminf_{N \to \infty} \frac{1}{N} \mathbb{E}_x^{\widetilde{\pi}} \left\{ \sum_{t=1}^{N} r(\widetilde{X_t}, \widetilde{Y_t}) \right\}. \tag{9}$$

The equivalence between $\mathcal{D}$ restricted MDP with finite state space S and the above defined MDP with a state space X of all probability distributions on S follows from the well-known equivalence between a partial observation MDP and an equivalent full observation MDP since we also have equivalence between $\mathcal{D}$ restricted MDP and a partial observation MDP as explained at the end of Section 2. Equivalence between a partial observation MDP and a corresponding full observation MDP is applied in [16] for the problem we mentioned before, while in [6] the equivalence is described and explored in a more general setup. Here we do not go in details about the equivalence between both models, but we note some useful aspects. Most important for this paper is the following property. Let $\widetilde{\pi}$ be a Markovian policy to be applied for the full observation MDP and $\widetilde{\omega} = (x_1, a_1, x_2, a_2, \ldots) \in \widetilde{\Omega}$ be an associated sample path. Define π as the $\mathcal{D}$-mixing policy defined by the infinite sequence of decision rules $(a_1, a_2, \ldots)$ corresponding to sample path $\widetilde{\omega}$. Then for the performances $g^{\widetilde{\pi}}(x_1)$ and $g^{\pi}(x_1)$ as defined by (9) and (1) respectively it almost surely holds that $g^{\widetilde{\pi}}(x_1) = g^{\pi}(x_1)$. Similarly it follows that if there exists an optimal stationary deterministic Markovian policy $\widetilde{\pi}$ for the full observation MDP then the deterministic $\mathcal{D}$ -mixing policy π obtained from the sample path $\widetilde{\omega}$ associated to $\widetilde{\pi}$ is an optimal $\mathcal{D}$-mixing policy for the initial state distribution

x_1.

Next two conditions for $\mathcal{D}$ restricted MDP are given. If they are satisfied then some useful results (Theorem **??** and Corollary 15) follow immdediately from the above explained equivalence between $\mathcal{D}$ restricted MDP and full observation MDP.

Condition 12 *For all deterministic $\mathcal{D}$-mixing policies π and initial state distributions $x, y \in X$ it holds that $g^\pi(x) = g^\pi(y)$. In other words performances of deterministic $\mathcal{D}$-mixing policies do not depend on the initial state distribution and the performance of such a policy π may be denoted by g^π.*

Condition 13 *There exist optimal stationary deterministic Markovian policies for the full observation MDP which is equivalent to the considered $\mathcal{D}$ restricted MDP.*

Theorem 14 *Suppose Condition 12 is satisfied for a $\mathcal{D}$ restricted MDP. Let $\widetilde{\pi}$ be a stationary deterministic Markovian policy for the equivalent full state MDP and $\widetilde{\omega} = (x_1, a_1, x_2, a_2, \ldots) \in \widetilde{\Omega}$ be an associated sample path. For $t = 1, 2, \ldots$ let π_t be the deterministic $\mathcal{D}$-mixing policy given by the infinite sequence of decision rules $(a_t, a_{t+1}, \ldots)$. Then the performances $g^{\widetilde{\pi}}(x_1)$ and $g^{\pi_t}(x_1)$ for $t = 1, 2, \ldots$ are independent of the initial state distribution x_1 and moreover, we have that*

$$g^{\widetilde{\pi}} = g^{\pi_t} \text{ for } t = 1, 2, \ldots. \tag{10}$$

Corollary 15 *Suppose Condition 12 and Condition 13 are satisfied for some $\mathcal{D}$ restricted MDP. Let $\widetilde{\pi}$ be an optimal stationary deterministic Markovian policy for the equivalent full state MDP and for $t = 1, 2, \ldots$ let π_t be the deterministic $\mathcal{D}$ mixing policy defined as in Theorem 14. Then for all $t = 1, 2, \ldots$ policy π_t is an optimal $\mathcal{D}$-mixing policy.*

5.1 On the existence of stationary optimal policies for the associated MDP

Next we wish to apply Corollary 15 to obtain structural results on optimal $\mathcal{D}$-mixing policies. However, to apply Corollary 15 Condition 12 and Condition 13 should be satisfied. In Example 11 we have seen that for Condition 12 to be satisfied it is not sufficient for the relevant transition matrices to be unichain and aperiodic. Moreover, since an equivalent full state MDP has an uncountable state space X it is in general also not obvious whether Condition 13 is satisfied. Therefore we wish to apply a sufficient condition according to Corollary 4.1 in [7] for Condition 13 to be satisfied. This result basically states that for an equivalent MDP with finite action set associated with a partially observable MDP with finite state space an uniformly boundedness condition is sufficient for the existence of an appropriate solution of the corresponding average cost optimality equation(ACOE) implying the existence of optimal stationary deterministic Markovian policies which applied to this paper means that Condition 13 is satisfied. For the cases in which we apply in this paper the uniformly boundedness condition to show that Condition 13 is satisfied it will follow that

Condition 12 is then also satisfied.
The uniformly boundedness condition as given in [7] is that the difference in optimal discounted costs is uniformly bounded over the state space X. Thus in [7] the conditions are stated with costs instead of rewards, but for rewards an analogous condition may be formulated. To formulate this analogous uniformly boundedness condition for rewards we first define the (optimal) discounted reward for the equivalent MDP with state space X.

Let $x \in X$ be an initial state and for $t = 1, 2, \ldots$ variables $\widetilde{X}_t$ and $\widetilde{Y}_t$ be defined as in (9) for policy $\widetilde{\pi}$. Since the number of components $|S|$ of both reward vectors $r(d^1), r(d^2)$ is finite with no loss of generality we assume in the sequel that all the components of the reward vectors are non-negative and bounded from above by some positive number B. Then for the obtained rewards $r(X_t, Y_t)$ of the reward process it follows that

$$0 \leq r(\widetilde{X}_t, \widetilde{Y}_t) \leq B \text{ for } t = 1, 2, \ldots. \tag{11}$$

We define for discount factor $0 < \beta < 1$, initial state $x \in X$ and policy $\widetilde{\pi}$ the discounted reward $R_\beta(x, \widetilde{\pi})$ (DR) by

$$R_\beta(x, \widetilde{\pi}) := \lim_{N \to \infty} \mathbb{E}_x^{\widetilde{\pi}} \left\{ \sum_{t=1}^{N} \beta^{t-1} r(\widetilde{X}_t, \widetilde{Y}_t) \right\}. \tag{12}$$

and the optimal β- discounted reward for initial state $x \in X$ is given by

$$R_\beta^*(x) := \sup_{\widetilde{\pi}} R_\beta(x, \widetilde{\pi}), \tag{13}$$

the supremum being taken over all admissible policies. By (11) it follows that the limit defining $R_\beta(x, \widetilde{\pi})$ exists and is finite and it is easily seen that $0 \leq R_\beta(x, \widetilde{\pi}) \leq \frac{B}{1-\beta}$ for all initial states $x \in X$ and admissible policies $\widetilde{\pi}$. From this it also follows for the optimal discounted reward that

$$0 \leq R_\beta^*(x) \leq \frac{B}{1-\beta}. \tag{14}$$

Since the action set $\widetilde{A} = \{d^1, d^2\}$ is finite and obtained rewards $r(\widetilde{X}_t, \widetilde{Y}_t)$ are bounded it follows (see for example [7]) that for every discount factor $0 < \beta < 1$ there exists some (finite) solution $R_\beta^*(x)$ satisfying the following discounted reward optimality equation (DROE):

$$R_\beta^*(x) = \max_{\widetilde{A} = \{d^1, d^2\}} \left\{ r(x, d^1) + \beta R_\beta^*(xP), r(x, d^2) + \beta R_\beta^*(xQ) \right\} \text{ for all } x \in X. \tag{15}$$

Moreover, by Theorem 4.2.3 in [14] it also follows that for every $0 < \beta < 1$ there exists some optimal policy which is Markovian, deterministic and stationary. Summarizing we have the following result.

Theorem 16 *For every discount factor $0 < \beta < 1$ there exists some solution $R^*_\beta(x)$ of (15) for which $0 \leq R^*_\beta(x) \leq \frac{B}{1-\beta}$ for all $x \in X$. Moreover, for the corresponding discounted MDP there exists some optimal policy which is Markovian, deterministic and stationary.*

The long-run expected average reward (AR) $g^{\widetilde{\pi}}(x)$ for policy $\widetilde{\pi}$ and initial state $x \in X$ was already given by (10) and similarly to the discounted case the optimal average reward for initial state $x \in X$ is given by

$$g^*(x) := \sup_{\widetilde{\pi}} g^{\widetilde{\pi}}(x) \tag{16}$$

A bounded solution of the average reward optimality equation (AROE) is a pair (g^*, h) satisfying

$$g^* + h(x) = \max_{\widetilde{A}=\{d^1,d^2\}} \left\{ r(x, d^1) + h(xP), r(x, d^2) + h(xQ) \right\} \text{ for all } x \in X \tag{17}$$

with $g^* \in \mathbb{R}$ and h a Borel measurable real-valued function on X, which is lower semi-continuous and bounded (i.e. $\sup_{x \in X} |h(x)| < \infty$).

If such a bounded solution (g^*, h) satisfying (17) exists then it also follows that there exists an optimal policy for the average reward criterion which is deterministic and stationair and moreover, $g^*(x) = g^*$ for any initial state $x \in X$. Thus the existence of such a bounded solution (g^*, h) of (17) is sufficient for Condition 13 to be satisfied and Condition 12 to be satisfied for (at least) all the optimal deterministic $\mathcal{D}$-mixing policies π which correspond to some sample path associated with an optimal deterministic stationary policy for the equivalent full state MDP. Now we give the uniformly boundedness condition for optimal discounted rewards which is sufficient for the existence of such a bounded solution of (17).

Condition 17 *There exists some $M \in \mathbb{R}$ such that for all $x, y \in X$ and $0 < \beta < 1$ it holds that*

$$|R^*_\beta(x) - R^*_\beta(y)| \leq M. \tag{18}$$

It turns out for Condition 17 to be satisfied that some sufficient conditions on transition matrices induced by decision rules in $\mathcal{D}$ may be formulated with help of Dobrushin's coefficient of ergodicity of a transition matrix.

Definition 18 *Let $P = (p_{ij})$ be a transition matrix on some finite state space S. Dobrushin's coefficient of ergodicity of P is defined as*

$$\rho_0(P) = \frac{1}{2} max_{i,j} \sum_{k=1}^{|S|} |p_{ik} - p_{jk}|. \tag{19}$$

Lemma 19 states some well-known (see for example [20]) useful properties of Dobrushin's coefficient.

Lemma 19

1. $0 \leq \rho_0(P) \leq 1$

2. $\rho_0(P) = 0$ *if and only if* P *has identical rows*

3. $\rho_0(P_1 \cdot P_2) \leq \rho_0(P_1) \cdot \rho_0(P_2)$

4. *There exists some positive integer* N *with* $\rho_0(P^N) < 1$ *if and only if* P *is unichain and aperiodic.*

For this paper the most useful property of Dobrushin's coefficient has to do with the l_1-distance between probability distributions on the finite state space S. For $x, y \in X$ denote by

$$||x - y||_1 := \sum_{i=1}^{|S|} |x_i - y_i|$$

the l_1-distance between probability distributions x and y on S. Then the following lemma (see [20]) holds.

Lemma 20 $||x - y||_1$ *is a metric on the set of probability distributions X with the property that $||x - y||_1 \leq 2$ for all $x, y \in X$. Moreover, for any $x, y \in X$ and transition matrix P on S we have that*

$$||xP - yP||_1 \leq \rho_0(P)||x - y||_1. \tag{20}$$

In other words if $\rho_0(P) < 1$ then P induces a contraction mapping on X. In the following results Lemma 19 and Lemma 20 will be applied to show that Condition 17 is satisfied under several specific assumptions on the transition matrices.

Theorem 21 *Let $\mathcal{D} = \{d^1, d^2\}$ and let P and Q be the transition matrix induced by decision rule d^1 respectively d^2. Assume that (at least) one of the two transition matrices has Dobrushin's coefficient smaller than 1 and both transition matrices are unichain and aperiodic. Then Condition 17 is satisfied.*

Proof. Let $x, y \in X$ be arbitrarily chosen probability distributions. We may assume without loss of generality that $R^*_\beta(x) \geq R^*_\beta(y)$ and then we should show that for all discount factors $0 < \beta < 1$ it holds that $R^*_\beta(x) - R^*_\beta(y) \leq M$ for some $M \in \mathbb{R}$. According to Theorem 16 for any arbitrarily chosen $0 < \beta < 1$ there exists some optimal Markovian deterministic and stationary policy $\widetilde{\pi}$. Let $\widetilde{\omega} = (x_1, a_1, x_2, a_2, \ldots)$ be the sample path with initial state $x_1 = x$ for optimal policy $\widetilde{\pi}$. Tracking sample path $\widetilde{\omega}$ for $t = 1, 2, \ldots$ let $r(a_t)$ be the reward vector for decision rule $a_t \in \mathcal{D}$, $A_t \in \{P, Q\}$ be the transition matrix corresponding to a_t and B_t be the matrix product given by $B_t := \prod_{k=1}^{t-1} A_t$ with the convention that B_1 is the identity matrix. Then $x_t = x_1 B_t$ for $t = 1, 2, \ldots$ and by (12) we have that

$$R^*_\beta(x) = R_\beta(x, \widetilde{\pi}) = \lim_{k\to\infty} \sum_{t=1}^{k} \beta^{t-1} r(x_t, a_t) = \lim_{k\to\infty} \sum_{t=1}^{k} \beta^{t-1} (xB_t) \cdot r(a_t). \tag{21}$$

The infinite sequence of decision rules $(a_1, a_2, \ldots)$ defines a policy $\widetilde{\pi}'$ for which the sample path $\widetilde{\omega}'$ for initial state $y_1 = y$ is given by $\widetilde{\omega}' = (y_1, a_1, y_2, a_2, \ldots)$ with $y_t = yB_t$ for $t = 1, 2, \ldots$. Hence

$$R^*_\beta(y) \geq R_\beta(y, \widetilde{\pi}') = \lim_{k\to\infty} \sum_{t=1}^{k} \beta^{t-1} r(y_t, a_t) = \lim_{k\to\infty} \sum_{t=1}^{k} \beta^{t-1} (yB_t) \cdot r(a_t) \tag{22}$$

Recall that we could assume that all components of the reward vectors $r(a_t)$ are nonnegative and bounded from above by some $B > 0$. Thus by (21),(22) and Lemma 20 we have

$$R^*_\beta(x) - R^*_\beta(y) \leq \lim_{k\to\infty} \sum_{t=1}^{k} \beta^{t-1} ((xB_t) \cdot r(a_t) - (yB_t) \cdot r(a_t)) =$$

$$\lim_{k\to\infty} \sum_{t=1}^{k} \beta^{t-1} (xB_t - yB_t) \cdot r(a_t) \leq \lim_{k\to\infty} \sum_{t=1}^{k} ||xB_t - yB_t||_1 B \leq$$

$$B \lim_{k\to\infty} \sum_{t=1}^{k} \rho_0(B_t) ||x - y||_1 \leq 2B \lim_{k\to\infty} \sum_{t=1}^{k} \rho_0(B_t). \tag{23}$$

Without loss of generality we may assume that $\rho_0(P) = \gamma_1 < 1$. Moreover, since Q is unichain and aperiodic there exists by property 4 of Lemma 19 some $N \in \mathbb{N}$ such that $\rho_0(Q^N) = \gamma_2 < 1$. Put $\gamma = \max(\gamma_1, \gamma_2)$. Then it follows by property 1 and property 3 of Lemma 19 that $\rho_0(B_{N+1}) \leq \gamma < 1$ since the matrix product B_{N+1} contains at least one P or $B_{N+1} = Q^N$. Similarly it follows that $\rho_0(B_{t+N}) \leq \gamma \rho_0(B_t)$ for $t = 1, 2, \ldots$. Combining this with $0 \leq \rho_0(B_t) \leq 1$ for $t = 1, 2, \ldots$ it follows that

$$\lim_{k\to\infty} \sum_{t=1}^{k} \rho_0(B_t) \leq N + \lim_{k\to\infty} \sum_{t=N+1}^{k} \rho_0(B_t) = N + \lim_{k\to\infty} \sum_{t=1}^{k} \rho_0(B_{t+N}) \leq N + \gamma \lim_{k\to\infty} \sum_{t=1}^{k} \rho_0(B_t).$$

Hence $(1-\gamma) \lim_{k\to\infty} \sum_{t=1}^{k} \rho_0(B_t) \leq N$ and thus $\lim_{k\to\infty} \sum_{t=1}^{k} \rho_0(B_t) \leq \frac{N}{1-\gamma}$. Combining this with (23) we obtain $R^*_\beta(x) - R^*_\beta(y) \leq \frac{2BN}{1-\gamma}$. Thus we have shown that for all $x, y \in X$ and $0 < \beta < 1$ it holds that

$$|R^*_\beta(x) - R^*_\beta(y)| \leq \frac{2BN}{1-\gamma} \tag{24}$$

Thus Condition 17 is satisfied with $M = \frac{2BN}{1-\gamma}$. ■

We have just shown that Condition 17 is satisfied if both transition matrices P and Q are unichain and aperiodic and at least one of them has Dobrushin coefficient smaller than one. Since Condition 17 is satisfied it also follows that Condition 13 is satisfied. Moreover, in the proof of Theorem 21 we have shown something additional which is also useful. Namely from the given proof it also follows that for any (not only an optimal) deterministic $\mathcal{D}$-mixing policy $\pi = (a_1, a_2, \ldots)$ and any time t the difference in expected accumulated total (undiscounted) rewards up to time t for any two initial state distributions $x, y \in X$ is uniformly bounded by $M = \frac{2BN}{1-\gamma}$. From this it immediately follows that the expected long-run average reward g^π of such a policy π does not depend on the initial state distribution. Thus Condition 12 is also satisfied for $\mathcal{D}$ restricted MDP as in Theorem 21. Thus Theorem 14 and Corollary 15 are applicable for such $\mathcal{D}$ restricted MDP. Consider the following example in which the results are applied.

Example 22 Consider once again the $\mathcal{D} = \{d^1, d^2\}$ restricted MDP considered in Example 7 and Example 9. Recall (6) describing P, Q, $r(d^1)$ and $r(d^2)$. It follows that $\rho_0(P) = 0.8$ and $\rho_0(Q) = 0.7$. Hence (24) holds for $B = 1$, $N = 1$ and $\gamma = 0.8$ and thus Condition 17 is satisfied for $M = 10$. Then as explained also Condition 13 and Condition 12 are satisfied. Hence Theorem 14 and Corollary 15 are applicable to obtain structural results on optimal $\mathcal{D}$ -mixing policies. Later we consider this example again to obtain such results.

Theorem 21 and its consequences may easily be generalized to be applicable for more $\mathcal{D}$ restricted MDP problems. Indeed from the proof it is easily seen that the conditions on the transition matrices given in Theorem 21 are a special case of the following more general result.

Theorem 23 *Consider a $\mathcal{D}$ restricted MDP with $\mathcal{D} = \{d^1, d^2, \ldots, d^n\}$ and let $\mathcal{A} = \{P_1, P_2, \ldots, P_n\}$ be the set of n corresponding transition matrices. Suppose there exists some $\gamma < 1$ and positive integer N such that for all n^N matrix products A of the form $A = \prod_{k=1}^{N} A_k$ with $A_k \in \mathcal{A}$ for $k = 1, 2, \ldots, N$ it holds that $\rho_0(A) \leq \gamma$, then for the equivalent (full observation) MDP Condition 17 is satisfied for $M = \frac{2BN}{1-\gamma}$. Moreover, Condition 13 and Condition 12 are also satisfied.*

Proof. Similar as in the proof of Theorem 21 it follows that (24) holds and thus Condition 17 is satisfied. Then as explained above it also follows that Condition 13 and Condition 12 are satisfied. ■

To conclude this section the following example gives for $\mathcal{D} = \{d^1, d^2\}$ a case where Theorem 23 is applicable while Theorem 21 is not applicable. Thus also for $\mathcal{D} = \{d^1, d^2\}$ Theorem 23 is really more general than Theorem 21.

Example 24 Consider the $\mathcal{D} = \{d^1, d^2\}$ restricted MDP with state space $S = \{1, 2, 3\}$. For decision rule d^1 the transition matrix P and reward vector $r(d^1)$ are as follows:

$$P = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 1 & 0 & 0 \\ 0.5 & 0.5 & 0 \end{pmatrix}, \quad r(d^1) = \begin{pmatrix} 2 \\ 0 \\ 3 \end{pmatrix}.$$

For the other decision rule d^2 the transition matrix Q and reward vector $r(d^2)$ are as follows:

$$Q = \begin{pmatrix} 1 & 0 & 0 \\ 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0.5 \end{pmatrix}, \quad r(d^2) = \begin{pmatrix} 0 \\ 2 \\ 0 \end{pmatrix}.$$

Then $\rho_0(P) = \rho_0(Q) = 1$ and thus Theorem 21 is not applicable in this case. However, it is easy to check that $\rho_0(P^2) = 0.75$, $\rho_0(Q^2) = 0.75$, $\rho_0(PQ) = 0.75$ and $\rho_0(QP) = 0.75$. Thus Theorem 23 is applicable with $N = 2$, $\gamma = 0.75$ and $B = 3$. Hence Condition 17 is satisfied for $M = 48$ and also Condition 13 and Condition 12 are satisfied. Thus Theorem 14 and Corollary 15 could also be applied in this case.

6 Optimizing over special subsets of deterministic mixing policies

To simplify notation and definitions we restrict again to the case that $\mathcal{D} = \{d^1, d^2\}$ which implies that deterministic mixing policies correspond to infinite sequences $U = (u_1, u_2, \ldots)$ of zeros and ones as explained before. However, we keep in mind that generalization to the case $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ could usually be achieved in a straightforward manner. The main issue for deterministic mixing policies is that optimization of the performance is not tractable if we consider the set W of all infinite sequences of zeros and ones. Let $W^p \subseteq W$ be the subset of all periodic sequences of zeros and ones. Assuming Condition 12 is satisfied for deterministic mixing policies corresponding to $U \in W^p$ we presented a formula (8) to compute the performance. The problem is that the set W^p is still both large and discrete and therefore it is not tractable to optimize the performance over all elements of W^p by enumeration. However, optimizing over specific relatively small subsets of W^p is tractable by enumeration of all performances. If the optimal performance within such a subset is close (or preferably even equal) to the optimal performance within W^p (and possibly also W) then we may obtain (almost) optimal mixing policies in such a way. Next we introduce some subsets for which optimization is more or less tractable and could give useful results.

At first it seems a good idea to optimize for some $n \in \mathbb{N}$ over the subset of periodic sequences with period smaller or equal than n. We denote such subset by $W^p(n) \subseteq W^p$. Optimization over such a subset may in practice give good results. For example consider the problem investigated in Example 7 and Example 9. Optimization over the small set $W^p(2)$ would result in the deterministic mixing policy corresponding to sequence $(1, 0)^\infty$ which according

to Example 9 yields a performance of 0.341 which improves the performance of the optimal Bernoulli mixing policy which equals 0.303 according to Example 7.
However, optimization over sets $W^p(n)$ has some disadvantages. First of all the cardinality of $W^p(n)$ increases exponentially in n and therefore it is only tractable for rather small n. Besides if n gets smaller than the optimal performance within the subset is likely to decrease. Thus there is a trade-off between computation time and performance and a priori it is unknown what would be a good choice for the maximal period n. For the problem of Example 7 we have seen that for period $n = 2$ already a policy exists which improves on the optimal Bernoulli policy, but for larger state space it is likely that a much larger period n is necessary to improve on the optimal Bernoulli policy. In general for fixed period n we can not say a priori whether the optimal performance over $W_p(n)$ is better than for the optimal Bernoulli mixing policy. Of course the optimal performance over $W_p(n)$ is not better than the optimal performance over W, but nothing is known about the difference. This lack of guarantees for the optimal performance over $W_p(n)$ motivates to investigate optimization over other kind of subsets.

Such an interesting other subset is the set G_0 of periodic sequences for which there exists some non-negative integer k such that between any two consecutive zeros there are exactly k other symbols in the sequence. In other words the gap between consecutive zeros is constant and equal to $k+1$. Since we restricted in this subsection to zero-one sequences this implies that there are exactly k ones between any two consecutive zeros and thus there exists a period cycle of the form $(0, 1, \ldots, 1)$. Analogously the set G_1 is the set of periodic zero-one sequences for which there exists some non-negative integer k such that between any two consecutive ones there are exactly k zeros in the sequence which implies that there exists a period cycle of the form $(1, 0, \ldots, 0)$. Note that the set $G := G_0 \cup G_1$ contains for example the only two decision sequences $0, 0, \ldots$ and $1, 1, \ldots$ inducing a stationary Markov chain. Moreover, we note that the periodic sequence with period cycle $(1, 0)$ is contained in both G_0 and G_1 and (up to cyclic shifts) is unique with respect to this property.

Optimization of the performance over the set G_0 (or G_1) corresponds to optimization over the non-negative integer k. Usually an optimal value for k is very small and easy to find. Then the optimal performance over G_0 (and also over G_1 or G) is quickly obtained. For example in Example 7 it may be shown that for $k = 1$ the performance over G_1 is optimized. For the corresponding policy with period cycle $(1, 0)$ we already computed the performance in Example 9. Recall that this performance 0.341 improved the optimal performance over all Bernoulli policies. Moreover, it may be checked that for Example 7 this sequence with performance 0.341 is also optimal over G_0 and thus over G. However, the question remains how close an optimal performance over G_0, G_1 or G is to the optimal performance over the much larger sets W_p or W. Some partial results are known on this which is more than for the earlier discussed $W_p(n)$ sets. Indeed for some particular problems it may be shown that a policy which corresponds to an optimal decision sequence over G is also optimal over W. For example in [16] the static assignment to parallel exponential servers with no buffer is studied. In [16] the problem to minimize the average number of blocked customers is formulated as stochastic control problem with partial observations. For 2 servers it is proved that

the optimal assignment sequence is in the set G. Moreover, if the server corresponding to symbol 1 is the faster server than the optimal assignment sequence is in G_0 and by symmetry it is in G_1 if the other server is faster. Note that from these results it also follows that if the two servers have equal service rates that there exists an optimal assignment sequence in $G_0 \cap G_1$ which implies that the round robin policy with period cycle $(1, 0)$ is optimal. In [1] similar results are obtained for general stationary arrival processes.

For $\mathcal{D} = \{d^1, d^2\}$ restricted MDP we have the following result on the existence of an optimal decision sequence within G_1 or G_0 in case the transition matrices satisfy some particular condition.

Theorem 25 *Consider a $\mathcal{D} = \{d^1, d^2\}$ restricted MDP. Let symbol 1 correspond to applying decision rule d^1 inducing an aperiodic and unichain transition matrix P. Let symbol 0 correspond to applying decision rule d^2 inducing an aperiodic and unichain transition matrix Q. If $\rho_0(P) = 0$ then there exists some optimal deterministic $\mathcal{D}$-mixing policy $\pi = (a_1, a_2, \ldots)$ for which the corresponding infinite decision sequence of zeros and ones is either $(0, 0, \ldots)$ or some element of G_1. Analogously if $\rho_0(Q) = 0$ then there exists some optimal deterministic $\mathcal{D}$-mixing policy $\pi = (a_1, a_2, \ldots)$ for which the corresponding infinite decision sequence of zeros and ones is either $(1, 1, \ldots)$ or some element of G_0.*

Proof. Suppose $\rho_0(P) = 0$ and Q is aperiodic and unichain. Then the conditions of Theorem 21 and thus also Theorem 23 are satisfied. Hence Condition 12 and Condition 13 are satisfied. Thus for the equivalent full observation MDP with state space X there exists some deterministic stationary Markovian policy $\widetilde{\pi}$ which is optimal. This deterministic stationary policy $\widetilde{\pi}$ may be represented by a mapping from state space X to action space $\widetilde{A} = \{d^1, d^2\}$. Let $a : X \to \widetilde{A}$ be this mapping. If policy $\widetilde{\pi}$ is applied then for any initial state distribution $x_1 \in X$ an associated sample path $\widetilde{\omega} = (x_1, a_1, x_2, a_2, \ldots)$ is inductively determined by

$$a_t = a(x_t) \text{ and } x_{t+1} = \begin{cases} x_t P & \text{if} \quad a_t = d^1 \\ x_t Q & \text{if} \quad a_t = d^2 \end{cases} . \tag{25}$$

For such $\widetilde{\omega}$ inductively determined by (25) it follows by Corollary 15 that for all $t \in \mathbb{N}$ the deterministic $\mathcal{D}$ mixing policy $\pi_t := (a_t, a_{t+1}, \ldots)$ is optimal with respect to maximizing the expected long-run average reward. Let $(u_1, u_2, \ldots)$ be the infinite sequence of zeros and ones corresponding to the infinite sequence of decision rules $(a_1, a_2, \ldots)$. Suppose $(u_1, u_2, \ldots)$ contains only a finite number of ones. Then there exists some $t \in \mathbb{N}$ such that $u_n = 0$ for all $t \geq n$ and it follows that the optimal deterministic $\mathcal{D}$ mixing policy π_t corresponds to the infinite decision sequence $(0, 0, \ldots)$. Suppose on the other hand that $(a_1, a_2, \ldots)$ contains infinitely many ones. Then there exist $k, l \in \mathbb{N}$ with $k < l$ such that $u_k = u_l = 1$ and $u_t = 0$ for all $k < t < l$. Since $\rho_0(P) = 0$ we have by Lemma 19 that P has identical rows and let $x_0 \in X$ be the unique row vector which is row of P. Then it follows that $xP = x_0$ for all $x \in X$. Since $a_k = a_l = d^1$ we have by (25) that $x_{k+1} = x_{l+1} = x_0$. From this it follows inductively by (25) that $x_t = x_{t+l-k}$ for all $t \geq k + 1$ and $a_t = a_{t+l-k}$ for all $t \geq k$. Thus the infinite sequence $(u_k, u_{k+1}, \ldots)$ is periodic with period $l - k$ and it follows that this sequence

is element of G_1 since we also have that $u_k = 1$ and $u_t = 0$ for all $k < t < l$. Thus we have proved Theorem 25 in case $\rho_0(P) = 0$. The statement for the case $\rho_0(Q) = 0$ follows analogously and a symmetry argument could be applied. ■

6.1 Regular sequences and corresponding policies

Consider again the $\mathcal{D} = \{d^1, d^2\}$ restricted MDP from Example 7 characterized by (6). Recall (Example 9) that the best performance we have obtained so far for this example is 0.341 which is obtained by the deterministic $\mathcal{D}$ mixing policy which corresponds to the periodic decision sequence with period cycle $(1, 0)$. Moreover, this is the best performance that can be obtained within the set G of decision sequences. Thus if $\rho_0(P) = 0$ or $\rho_0(Q) = 0$ then it would follow from Theorem 25 that this policy with performance 0.341 would be optimal over all $\mathcal{D}$ mixing policies, but $\rho_0(P) = 0.8 > 0$ and $\rho_0(Q) = 0.7 > 0$. Thus in this example it could be possible that some decision sequence which is not an element of G has better performance. Indeed by (8) it can be checked that for example the periodic decision sequence with period cycle $(1, 1, 0, 1, 0, 1, 0, 1, 0)$ yields an expected long-run average reward of 0.3435 which slightly improves 0.341. Thus in this example there exist periodic decision sequences with better performance than the optimal performance within the set G.

The remaining question is whether such improving sequences can only be found by an exhaustive search over the set W (or W_p) or if it is possible to characterize some subset of W in which improving decision sequences can be found provided they exist. Such characterization is useful if searching over the subset is less complex than over W (W_p) and then it would be especially nice if the optimal decision sequence is shown to be element of this subset. Indeed we may characterize some subset of W which potentially has all these desired properties. This is the subset $\mathcal{R} \subseteq W$ of so-called regular sequences of zeros and ones. In the sequel of this paper we define this subset and give some of the most useful properties and characterizations. We show how an effective optimization over this subset $\mathcal{R}$ may be performed and we will apply this to the $\mathcal{D} = \{d^1, d^2\}$ restricted MDP from Example 7. Finally we give some conditions which are shown to be sufficient that some optimal $\mathcal{D}$-mixing policy corresponds to a decision sequence which is a regular sequence. Thus in that case the optimal performance may indeed be found within this set of regular sequences.

Definition 26 *Let $U = (u_1, u_2, \ldots)$ be an infinite sequence of certain symbols. A suffix of U is an infinite sequence of the form $(u_n, u_{n+1}, \ldots)$ for some $n \in \mathbb{N}$. A finite subsequence of U is a finite sequence of the form $(u_k, u_{k+1}, \ldots, u_l)$ for some $k, l \in \mathbb{N}$ with $k \leq l$.*

In the sequel $U = (u_1, u_2, \ldots)$ is assumed to be an infinite sequence of zeros and ones. In that case we denote by $s_k(n) := \sum_{j=k}^{k+n-1} u_j$ the number of ones in the subsequence of length n beginning at the k-th element of U and put $s(n) := s_1(n)$. Then U is said to have a density of $\theta \in [0, 1]$ if $\lim_{n\to\infty} \frac{s(n)}{n} = \theta$. Thus if an infinite sequence U of zeros and ones has a density θ then θ is the asymptotic frequency of the ones in U. In that case it may intuitively be clear

that the positions of ones in the sequence are more regularly distributed if for all $k, n \in \mathbb{N}$ absolute deviations between $s_k(n)$ and $n\theta$ are small. The following fundamental definition is based on this intuition and defines exactly when an infinite sequence of zeros and ones is (most) regular. We also define when a sequence is so-called eventually regular.

Definition 27 *Let $U = (u_1, u_2, \ldots)$ be an infinite sequence of zeros and ones. Then U is called regular of density θ if for $s_k(n)$, the number of ones in the corresponding subsequence of length n, it holds that*

$$|s_k(n) - n\theta| < 1 \text{ for every } k, n \in \mathbb{N}. \tag{26}$$

An infinite sequence of zeros and ones is called eventually regular if it has a suffix which is regular of some density.

The earlier discussed subset $\mathcal{R} \subseteq W$ of regular sequences can now be defined as the set of infinite sequences of zeros and ones which are regular for some density $\theta \in [0, 1]$. It is obvious that if some sequence U is regular and thus element of $\mathcal{R}$ that (26) holds for some unique $\theta \in [0, 1]$ which will be the density of the sequence. For infinite sequences of zeros and ones very closely related to this notion of being (eventually) regular but possibly more convenient to apply is the notion of being (eventually) balanced. This notion is defined as follows.

Definition 28 *Let $U = (u_1, u_2, \ldots)$ be an infinite sequence of zeros and ones. Then U is called balanced if*

$$|s_k(n) - s_l(n)| \leq 1 \text{ for every } k, l, n \in \mathbb{N}. \tag{27}$$

In other words U is balanced if for any two finite subsequences of the same length the number of ones contained in these subsequences differs by at most one.
An infinite sequence of zeros and ones is called eventually balanced if it has a suffix which is balanced.

A complete classification of balanced sequences was given in [19]. Next we enumerate in Proposition 29 and Proposition 30 for regular (balanced) sequences the most important properties and connections which are useful for the present paper. These results are obvious or may be retrieved from results in [19], [23] or [17] although in these references not exactly the same terminology is used.

Proposition 29 *For infinite sequences of zeros and ones the following properties hold.*

1. *All regular sequences are balanced.*
2. *All balanced sequences are eventually regular.*
3. *A sequence is eventually regular if and only if it is eventually balanced.*
4. *Let $U = (u_1, u_2, \ldots)$ be an infinite sequence of zeros and ones and $V = (v_1, v_2, \ldots)$ be defined by $v_n = 1 - u_n$ for all $n \in \mathbb{N}$. Then U is balanced if and only if V is balanced. Moreover, U is regular of density θ if and only if V is regular of density $1 - \theta$.*

5. *For every $\theta \in [0,1]$ there exist some regular sequence(s) of density θ. Indeed for given $\theta \in [0,1]$ a regular sequence $U = (u_1, u_2, \ldots)$ of density θ may be obtained as follows. Choose some arbitrary $\phi \in \mathbb{R}$ and let U be determined either by*

$$u_n = \lfloor n\theta + \phi \rfloor - \lfloor (n-1)\theta + \phi \rfloor \text{ for all } n \in \mathbb{N} \tag{28}$$

or by

$$u_n = \lceil n\theta + \phi \rceil - \lceil (n-1)\theta + \phi \rceil \text{ for all } n \in \mathbb{N}. \tag{29}$$

Then U is regular of density θ. Moreover, an infinite sequence of zeros and ones $(u_1, u_2, \ldots)$ may be determined for some $\phi \in \mathbb{R}$ by either (28) or (29) if and only if the sequence is regular of density θ.

6. *A regular sequence of density θ is periodic if and only if θ is rational. If $\theta = \frac{p}{q}$ with $p, q \in \mathbb{N}$, p and q coprime, then the regular sequence has a period cycle of length q containing exactly p ones and $q - p$ zeros.*

Also noteable is that the set G introduced in the previous section is a subset of $\mathcal{R}$. Indeed it is easily seen that G_0 corresponds to regular sequences of densities $0, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \ldots$ and G_1 corresponds to regular sequences of densities $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots$. Thus the maximal performance over $\mathcal{R}$ is always at least as good as the maximal performance over G. The following result will be very useful for implementing the maximization of performance over the set $\mathcal{R}$.

Proposition 30 *Let U and V be regular sequences and suppose they both have density θ. Then the set of all finite subsequences of U equals the set of all finite subsequences of V. Moreover, either V is a suffix of U or U is a suffix of V. If θ is rational then the period cycles of U and V are cyclic shifts of each other.*

For example $U = (1,0,1,0,0)^\infty$ and $V = (0,1,0,1,0)^\infty$ are regular sequences of the same density $\frac{2}{5}$ and indeed the period cycles $(1,0,1,0,0)$ and $(0,1,0,1,0)$ are cyclic shifts of each other.

6.2 Optimization over regular sequences

We have defined the subset $\mathcal{R}$ of regular sequences and described some important properties of regular (balanced) sequences. In this subsection our objective is to apply this and optimize the performance over $\mathcal{R}$ in an efficient manner. Regular and/or balanced sequences have been applied in open-loop control of particular queueing systems. In [9] it was proved for some specific admission control problem that the optimal control sequence is a regular sequence. After that these sequences have been applied (see for example [4], [2], [3], [15] and [8]) to several admission, routing and polling problems. In such applications to queueing and discrete-event systems the optimality of regular sequences for open-loop control follows from multimodularity of an appropriate performance criterion such as expected workload in a queue or expected waiting times. Multimodularity is a property of functions defined on a

discrete set which is comparable to convexity for functions defined on a continuous set. The concept of multimodularity and its applications are discussed in detail in [1] and then also an overview of control problems is given for which optimality of regular sequences can be established by multimodularity. In [1] several assumptions like for example specifications on the topology of the queueing system are used to obtain multimodularity.

In the present paper the objective is to apply regular sequences for general $\mathcal{D}$ restricted MDP optimizing the long-run average reward instead of some specific open-loop queueing control problem with a specific performance criterion like for example minimizing expected average waiting times of customers as in previous papers. A consequence of this that multimodularity of the performance function is not applicable to get results on the optimality of some policy corresponding to a regular decision sequence. In the next section we show that if for the equivalent full observation MDP an optimal stationary and deterministic policy exists satisfying some specific properties that the existence of an optimal $\mathcal{D}$ mixing policy corresponding to some regular sequence follows. This is a new approach to establish without multimodularity the optimality of regular sequences for some (restricted) MDP problems. In this subsection we discuss and illustrate with an example the remaining practical problem of optimizing the performance over $\mathcal{R}$, the set of all regular sequences of zeros and ones.

When we optimize over $\mathcal{R}$ for some $\mathcal{D} = \{d^1, d^2\}$ restricted MDP problem we assume that Condition 12 and Condition 13 are satisfied such that Theorem 14 and Corollary 15 are applicable. In Section 5 we have seen some sufficient conditions for this. Then by Theorem 14 and Proposition 30 it follows that all deterministic $\mathcal{D}$ mixing policies corresponding to regular sequences of the same density $\theta \in [0, 1]$ have the same performance. Thus we may denote by $h(\theta)$ the long-run average reward of a deterministic $\mathcal{D}$ mixing policy corresponding to a regular decision sequence of density θ. Now we have that maximizing the performance over $\mathcal{R}$ is nothing more than maximizing the function $h(\theta)$ over $\theta \in [0, 1]$. Recall from Section 3 that this problem is rather similar to finding the optimal Bernoulli policy for which a performance function $g(\theta)$ should be maximized over $\theta \in [0, 1]$. We also note that in the admission, routing and polling problems in which regular sequences have been applied before the optimization in most cases was reduced to a maximization or minimization over the density θ of the regular sequence. Then it always held that $h(\theta) \geq g(\theta)$ for all θ in case of maximization or $h(\theta) \leq g(\theta)$ in case of minimization. Hence the optimal value of $h(\theta)$ improves the optimal performance over all Bernoulli policies. Naturally we expect and would like that this property also holds for $\mathcal{D}$ restricted MDP problems like we consider in the present paper.

Recall that for Bernoulli policies it is not difficult to maximize $g(\theta)$ since for any $\theta \in [0, 1]$ the value $g(\theta)$ can be computed quickly and possibly a closed formula for $g(\theta)$ may be obtained as in Example 7. However, maximizing $h(\theta)$ is more difficult. First of all it seems in general impossible to obtain a closed formula for $h(\theta)$ and for irrational θ we do not even have a finite algorithm to compute $h(\theta)$. On the other hand if θ is rational the value $h(\theta)$ may be computed by the following finite algorithm.

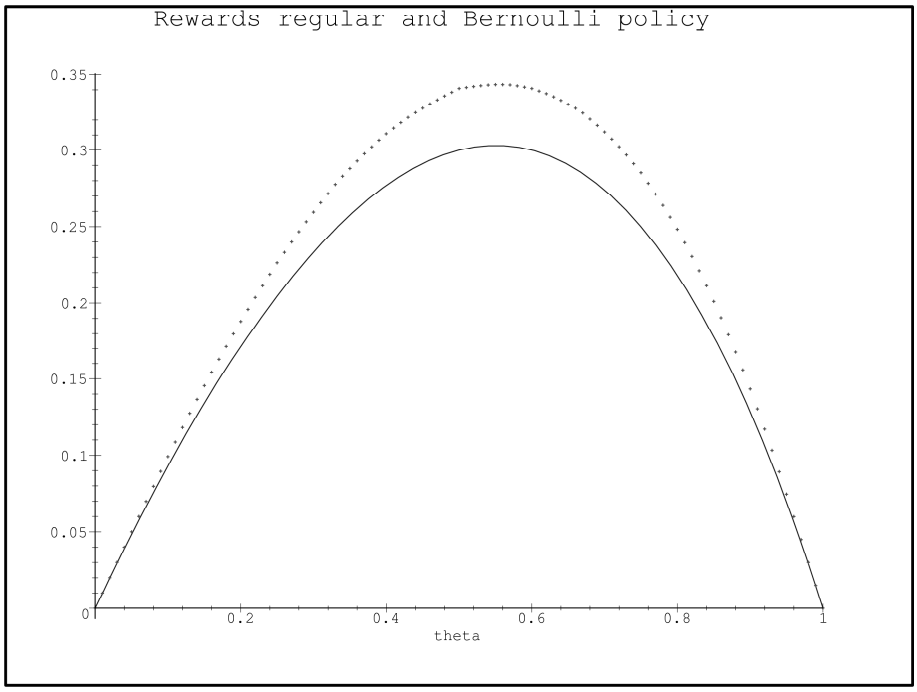

Figure 1: The performance of regular sequences versus Bernoulli policies

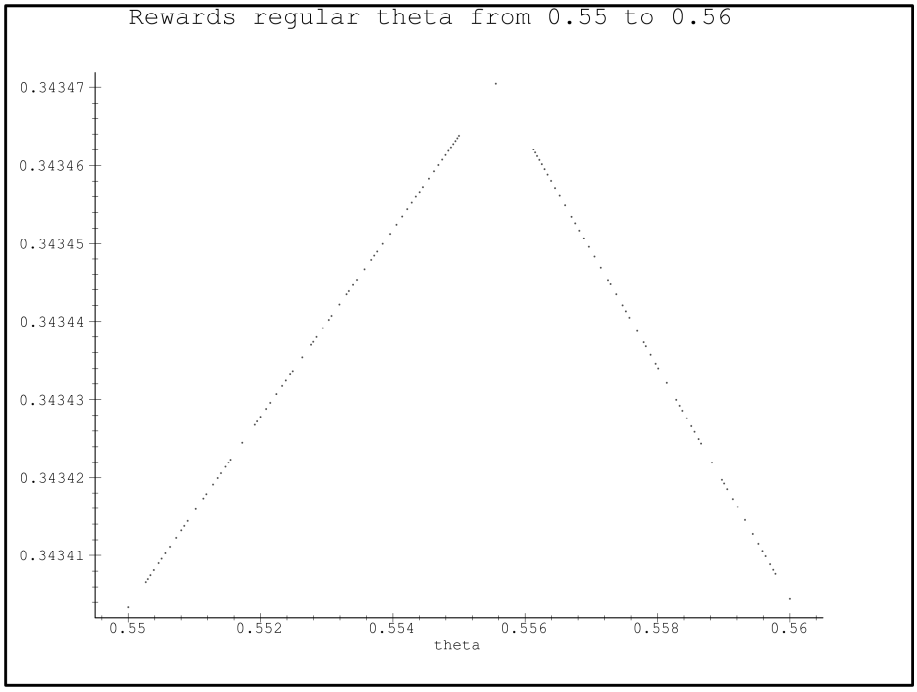

Figure 2: The performance of regular sequences for rational densities in the interval [0.55,0.56] and denominator at most 200

Algorithm 31 *If Condition 12 is satisfied this algorithm computes the performance $h(\theta)$ of any $\mathcal{D} = \{d^1, d^2\}$ mixing policy corresponding to a regular decision sequence of rational density $\theta \in [0,1]$.*

1. *Determine coprime integers p and q with $p \geq 0$, $q > 0$ such that $\theta = \frac{p}{q}$.*

2. *Choose some default value for ϕ, say $\phi = 0$, and then for $n = 1, 2, \ldots, q$ compute u_n by (28). The obtained sequence $(u_1, u_2, \ldots, u_q)$ is a period cycle of a regular sequence of density $\theta = \frac{p}{q}$.*

3. *Apply (8) to compute the long-run average reward g^π of the periodic policy π with period cycle $(u_1, u_2, \ldots, u_q)$. The value $h(\theta)$ is obtained by putting $h(\theta) = g^\pi$.*

The running time of Algorithm 31 increases in the denominator q of θ since the period cycle of the regular sequence of density θ is of length q. Thus for given $\theta = \frac{p}{q}$ the computation time is of order $\Omega(q)$ and to obtain or approximate the maximal value of $h(\theta)$ it seems most efficient to apply Algorithm 31 to a set of densities θ with bounded denominator q. For example Algorithm 31 can be applied to obtain a maximum of $h(\theta)$ over the set $\mathcal{R} \cap W_p(n)$ for some $n \in \mathbb{N}$. For such maximization the algorithm has to be applied only $O(n^2)$ times and for each run the period cycle of the decision sequence is at most n. Therefore the total computation time is polynomial in n and the algorithm terminates relatively quickly if neither n nor the state space are very large.
For the $\mathcal{D}$ restricted MDP from Example 7 the algorithm quickly maximizes the performance over $\mathcal{R} \cap W_p(n)$ for a maximal period of for example $n = 200$. Applying the algorithm it follows that the regular sequence with period cycle $(1,1,0,1,0,1,0,1,0)$ and density $\theta = \frac{5}{9}$ maximizes the performance over this set. Recall from the previous subsection that this particular decision sequence yields an expected long-run average reward of 0.3435. Applying the Algorithm for larger values of n does not give another improvement. Results in the sequel of this paper support the optimality of this regular sequence of density $\frac{5}{9}$ for the $\mathcal{D}$ restricted MDP from Example 7.

We note that that $\frac{5}{9}$ is close but not equal to $\theta^* = 3 - \sqrt{6} \sim 0.551$ which maximizes (recall Example 7) the performance over Bernoulli policies over rate θ. Figure 1, in which for $\theta \in [0,1]$ the performance of Bernoulli policies and deterministic $\mathcal{D}$-mixing policies given by a regular sequence of density θ are plotted, illustrates this. Recall from Example 7 that for Bernoulli policies of rate θ the performance $g(\theta)$ is according to the function $g(\theta) = \frac{3\theta - 3\theta^2}{3-\theta}$. For regular sequences of density θ the performance $h(\theta)$ is plotted for all $\theta = \frac{k}{100}$ for $k = 0, 1, \ldots, 100$. Thus $g(\theta)$ is the solid smooth curve in Figure 1, while the isolated points $(\theta, h(\theta))$ for $\theta = \frac{k}{100}$ also seem to be situated on some smooth curve. This suggests that the performance $h(\theta)$ for regular sequences is continuous for $\theta \in [0,1]$ just as $g(\theta)$ which we proved to be continuous. It is also interesting to note that $h(\theta)$ is never smaller than $g(\theta)$ and that the difference in performance $h(\theta) - g(\theta)$ appears to be maximal around the value where $h(\theta)$ is maximal. Moreover, Figure 1 confirms that the value of θ which maximizes $h(\theta)$ could well be $\theta = \frac{5}{9}$ and that the maximizing value for $h(\theta)$ is close to the value which

maximizes $g(\theta)$.
Figure 2 visually confirms that for $\theta = \frac{5}{9}$ the value of $h(\theta)$ is maximal. In Figure 2 the value of θ is varying over the small interval $[0.55, 0.56]$ and in this interval all points $(\theta, h(\theta))$ are plotted for all rational $\theta = \frac{m}{n}$ with denominator $n \leq 200$. In this figure the point $(\frac{5}{9}, h(\frac{5}{9}))$ is obviously the top one. Moreover, from the triangular shape which is recognizable in Figure 2 it may be concluded that for θ in this small interval around $\frac{5}{9}$ the value of $h(\theta)$ increases approximately linearly if θ approximates $\frac{5}{9}$.

7 Sufficient conditions for optimality of a regular sequence

In this final section we show that certain conditions for $\mathcal{D} = \{d^1, d^2\}$ restricted MDP are sufficient for the existence of an optimal $\mathcal{D}$-mixing policy which is deterministic corresponding to a *regular* zero-one decision sequence. We also discuss the applicability of the results to $\mathcal{D}$ restricted MDP problems and in particular the problem introduced in Example 7.
First we formulate and prove a key result which states that some infinite sequence of zeros and ones generated by iterating some function on the interval $[0, 1]$ is eventually regular if the functions satisfies certain conditions. In the sequel we denote with I the interval $[0, 1]$.

Iteration 32 *Let $x_1, x^* \in I$ be given. Let $f_1, f_2 : I \to I$ be given functions and $f : I \to I$ be defined by*

$$f(x) = \begin{cases} f_1(x) & \text{if } x \leq x^* \\ f_2(x) & \text{if } x > x^* \end{cases} \tag{30}$$

Consecutively for $n = 1, 2, \ldots$ determine u_n and x_{n+1} iteratively by

$$u_n := \begin{cases} 0 & \text{if } x_n \leq x^* \\ 1 & \text{if } x_n > x^* \end{cases} \quad \text{and } x_{n+1} := f(x_n). \tag{31}$$

Theorem 33 *Let $U = (u_1, u_2, \ldots)$ be an infinite sequence of zeros and ones generated by Iteration 32 with $f_1, f_2 : I \to I$ satisfying the following properties:*

1. *f_1 and f_2 are monotonically increasing.*
2. $$f_1(f_2(x)) \geq f_2(f_1(x)) \text{ for all } x \in I. \tag{32}$$

Then U is an eventually regular sequence.

To prove Theorem 33 we apply Lemma 34 which follows immediately from Proposition 2.1.3 in [17]. Similar as in [17] for sequences (or so-called words) $a = (a_1, a_2, \ldots, a_n)$ and $b = (b_1, b_2, \ldots, b_m)$ the concatenation $(a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_m)$ is shortly denoted with ab.

Lemma 34 *Let $U = (u_1, u_2, \ldots)$ be an infinite sequence of zeros and ones. Then U is balanced if and only if there does not exist some (possibly empty) finite sequence w of zeros and ones such that both $0w0$ and $1w1$ are subsequences of U.*

Proof of Theorem 33. We distinguish a few cases. In the first case suppose $f_1(x^*) \leq x^*$. Then $f_1(x) \leq x^*$ for all $0 \leq x \leq x^*$ since f_1 is monotonically increasing. Thus if $x_n \leq x^*$ for some $N \in \mathbb{N}$ then $x_n \leq x^*$ for $n = N, N+1, \ldots$ and thus $(u_N, u_{N+1}, \ldots) = (0, 0, \ldots)$ is regular of density 0. Hence U is an eventually regular sequence. If on the other hand $x_n > x^*$ for $n = 1, 2, \ldots$ then $U = (1, 1, \ldots)$ is regular of density 1.
In the second case suppose $f_2(x^*) > x^*$. Then it follows analogously to the first case that there exist some $N \in \mathbb{N}$ such that $x_n > x^*$ for all $n \geq N$ or $x_n \leq x^*$ for $n = 0, 1, \ldots$. Hence either U is eventually regular of density 1 or U is regular of density 0.
In the third and most important case we suppose that $f_2(x^*) \leq x^* < f_1(x^*)$ and let J denote the interval $[f_2(x^*), f_1(x^*)]$. Note that if $x_n < f_2(x^*)$ then $x_{n+1} = f_1(x_n) \leq f_1(x^*)$ and if $x_n > f_1(x^*)$ then $x_{n+1} = f_2(x_n) \geq f_2(x^*)$. Thus either $x_n < f_2(x^*) \leq x^*$ for $n = 1, 2, \ldots$ or $x_n > f_1(x^*) > x^*$ for $n = 1, 2, \ldots$ or $x_N \in J$ for some $N \in \mathbb{N}$. Thus either U is regular of density 0 or U is regular of density 1 or $x_N \in J$ for some $N \in \mathbb{N}$. Suppose $x_N \in J$ for some $N \in \mathbb{N}$. If $f_2(x^*) \leq x_N \leq x^*$ then $x_{N+1} = f_1(x_N) \leq f_1(x^*)$ and by (32) we also have that

$$x_{N+1} = f_1(x_N) \geq f_1(f_2(x^*)) \geq f_2(f_1(x^*)) \geq f_2(x^*)$$

and thus $x_{N+1} \in J$. Similarly if $x^* < x_N \leq f_1(x^*)$ then $x_{N+1} = f_2(x_N) \geq f_2(x^*)$ and by (32) we also have that

$$x_{N+1} = f_2(x_N) \leq f_2(f_1(x^*)) \leq f_1(f_2(x^*)) \leq f_1(x^*)$$

and thus $x_{N+1} \in J$. Hence if $x_N \in J$ then it follows by induction that $x_n \in J$ for $n = N, N+1, \ldots$. Thus in this third case we may assume that there exist some $N \in \mathbb{N}$ such that $x_n \in J$ for all $n \geq N$.
Consider the suffix $U' := (u_N, u_{N+1}, \ldots)$ of U. Suppose $u_m = 0$, $u_n = 1$ for some $m, n \in N$. Assume there exists some $k \in \mathbb{N}$ for which $u_{m+k} \neq u_{n+k}$ and let k_0 be the minimal positive integer satisfying $u_{m+k_0} \neq u_{n+k_0}$. We claim that then it follows that

$$u_{m+k_0} = 1 \text{ and } u_{n+k_0} = 0.$$

To verify this claim note that by (32) and $x_n, x_m \in J$ we have that

$$x_{m+1} = f_1(x_m) \geq f_1(f_2(x^*)) \geq f_2(f_1(x^*)) \geq f_2(x_n) = x_{n+1}.$$

Thus if $k_0 = 1$ then $x_{m+k_0} \geq x_{n+k_0}$ implying $1 \geq u_{m+k_0} \geq u_{n+k_0} \geq 0$. Hence $u_{m+k_0} = 1$ and $u_{n+k_0} = 0$ follows from the fact that $u_{m+k_0} \neq u_{n+k_0}$. If $k_0 \geq 2$ then we have that $u_{m+1} = u_{n+1}$. Since both f_1, f_2 are monotonically increasing and thus order-preserving it follows that $x_{m+2} \geq x_{n+2}$ by either $x_{m+2} = f_1(x_{m+1}) \geq f_1(x_{n+1}) = x_{n+2}$ or $x_{m+2} = f_2(x_{m+1}) \geq f_2(x_{n+1}) = x_{n+2}$. By applying this order preserving property of both f_1 and f_2 repetively it follows again that $x_{m+k_0} \geq x_{n+k_0}$ and thus $u_{m+k_0} = 1$ and $u_{n+k_0} = 0$ as above. Thus the claim holds but then it follows that there does not exist some (possibly empty) finite sequence w of zeros and ones such that both $0w0$ and $1w1$ are subsequences

of U'. Thus by Lemma 34 it follows that U' is balanced. Thus by definition U is eventually balanced since U' is a suffix of U and by Proposition 29 it follows that U is an eventually regular sequence. ■

Next our aim is to apply Theorem 33 to $\mathcal{D} = \{d^1, d^2\}$ restricted MDP problems with finite state space S satisfying some specific properties. For this we consider again the equivalent full observation MDP with continuous state space X of probability distributions on S as we introduced in Section 5. First of all we restrict to problems for which Condition 12 and Condition 13 are satisfied. Recall that in Subsection 5.1 we have investigated when these two conditions are satisfied and that we have seen that they are satisfied for a considerable class of problems. Now we define an extra condition which should hold in particular for the applicability of Theorem 33. In the sequel this new condition will be called the threshold condition since basically it says that for the equivalent full observation MDP some optimal stationary deterministic Markovian policy (which exists according to Condition 13) has some "threshold structure". In Definition 35 we define this notion of "threshold structure" for such policies which is followed by Condition 36 stating our threshold condition for $\mathcal{D} = \{d^1, d^2\}$ restricted MDP.

Definition 35 *Let $h : X \to \mathcal{A} = \{d^1, d^2\}$ be the mapping corresponding to a stationary deterministic Markovian policy $\widetilde{\pi}$. Then we say that mapping h and policy $\widetilde{\pi}$ have threshold structure if there exists some $i \in S$ and $x^0 \in I$ such that for all $x = (x_1, x_2, \ldots, x_{|S|}) \in X$ we either have that $h(x) = d^1$ if and only if $x_i \leq x^0$ ($x_i < x^0$) or $h(x) = d^2$ if and only if $x_i \leq x^0$ ($x_i < x^0$).*

Condition 36 *For the full observation MDP which is equivalent to the considered $\mathcal{D} = \{d^1, d^2\}$ restricted MDP there exist some optimal stationary deterministic Markovian policy $\widetilde{\pi}$ having a threshold structure as defined in Definition 35.*

Proposition 37 connects Condition 36 with Iteration 32 in case of a 2 states state space as for example in the $\mathcal{D} = \{d^1, d^2\}$ restricted MDP of Example 7. Then Theorem 33 will be applicable if the appropriate functions f_1 and f_2 have the properties stated in Theorem 33. Additionally from Proposition 37 it follows for such 2 state cases the appropriate f_1 and f_2 are linear which in the sequel will be useful to check the properties to apply Theorem 33.

Proposition 37 *Consider a $\mathcal{D} = \{d^1, d^2\}$ restricted MDP with state space $S = \{1, 2\}$. Suppose that Condition 36 is satisfied and let $\widetilde{\pi}$ be a stationary deterministic Markovian policy for the equivalent full observation MDP having a threshold structure. Let $\widetilde{\omega} = (y_1, a_1, y_2, a_2, \ldots) \in \widetilde{\Omega}$ be an associated sample path. For $n = 1, 2, \ldots$ let $v_n, w_n \in I$ satisfying $v_n + w_n = 1$ be such that $y_n = (v_n, w_n)$. Then there exist $x_1, x^* \in I$ and linear functions $f_1, f_2 : I \to I$ such that the sequences $(u_1, u_2, \ldots)$ and $(x_1, x_2, \ldots)$ generated by Iteration 32 satisfy the following properties.*

1. *Either $x_n = v_n$ for $n = 1, 2, \ldots$ or $x_n = w_n$ for $n = 1, 2, \ldots$.*

2. Either (33) or (34) holds,

$$u_n = \begin{cases} 0 & \text{if } a_n = d^1 \\ 1 & \text{if } a_n = d^2 \end{cases} \quad \text{for } n = 1, 2, \ldots \tag{33}$$

$$u_n = \begin{cases} 0 & \text{if } a_n = d^2 \\ 1 & \text{if } a_n = d^1 \end{cases} \quad \text{for } n = 1, 2, \ldots. \tag{34}$$

Proof. Let P be the transition matrix corresponding to d^1 and Q be the transition matrix corresponding to d^2. Let $a, b, c, d \in I$ be such that

$$P = \begin{pmatrix} a & 1-a \\ 1-b & b \end{pmatrix}, \quad Q = \begin{pmatrix} c & 1-c \\ 1-d & d \end{pmatrix}. \tag{35}$$

Let $h : X \to \{d^1, d^2\}$ be the mapping corresponding to $\widetilde{\pi}$. Then h has a threshold structure (see Definition 35) and assume h has this property for state $i = 1$. Now we distinguish several cases.

In the first case suppose that there exists some $x^0 \in I$ such that for any $\hat{x} = (\hat{x}_1, \hat{x}_2) \in X$ it holds that $h(\hat{x}) = d^1$ if and only if $\hat{x}_1 \leq x^0$. Then we claim that by putting $x_1 = v_1$, $x^* = x^0$, $f_1(x) = (a + b - 1)x + 1 - b$ for all $x \in I$ and $f_2(x) = (c + d - 1)x + 1 - d$ for all $x \in I$ the sequences $(u_1, u_2, \ldots)$ and $(x_1, x_2, \ldots)$ generated by Iteration 32 satisfy $x_n = v_n$ for $n = 1, 2, \ldots$ and moreover, $u_n = 0$ if and only if $a_n = d^1$. We prove this claim by induction to n. For $n = 1$ we already have $x_1 = v_1$. If $v_1 \leq x^0$ then $a_1 = h(y_1) = d^1$, $x_1 \leq x^*$ and thus $u_1 = 0$. On the other hand if $v_1 > x^0$ then $a_1 = h(y_1) = d^2$, $x_1 > x^*$ and thus $u_1 = 1$. Thus the claim holds for $n = 1$. Suppose the claim holds for $n = k$ and thus $x_k = v_k$. Distinguish the cases $v_k \leq x^0$ and $v_k > x^0$. Suppose $v_k \leq x^0$. Then $a_k = h(y_k) = d^1$ and thus $u_k = 0$ by the induction claim. Then by Iteration 32 it follows that $x_{k+1} = f_1(x_k) = f_1(v_k) = (a + b - 1)v_k + 1 - b$. Also we have

$$y_{k+1} = (v_k, w_k)P = (v_k, w_k)\begin{pmatrix} a & 1-a \\ 1-b & b \end{pmatrix} = (av_k + (1-b)w_k, (1-a)v_k + bw_k).$$

Hence $v_{k+1} = av_k + (1-b)w_k = av_k + (1-b)(1-v_k) = (a + b - 1)v_k + 1 - b$ and thus $x_{k+1} = v_{k+1}$ if $v_k \leq x^0$.
Suppose $v_k > x^0$. Then $a_k = h(y_k) = d^2$ and thus $u_k = 1$ by the induction claim. Then by Iteration 32 it follows that $x_{k+1} = f_2(x_k) = f_2(v_k) = (c + d - 1)v_k + 1 - d$. Also we have

$$y_{k+1} = (v_k, w_k)Q = (v_k, w_k)\begin{pmatrix} c & 1-c \\ 1-d & d \end{pmatrix} = (cv_k + (1-d)w_k, (1-c)v_k + dw_k).$$

Hence $v_{k+1} = cv_k + (1-d)w_k = cv_k + (1-d)(1-v_k) = (c + d - 1)v_k + 1 - d$ and thus $x_{k+1} = v_{k+1}$ if $v_k > x^0$. Thus we have proved that $x_{k+1} = v_{k+1}$. Then for $n = k + 1$ it follows that $u_n = 0$ if and only if $a_n = d^1$ similarly as for $n = 1$ and the induction proof is finished.

In the second case suppose that there exists some $x^0 \in I$ such that for any $\hat{x} = (\hat{x}_1, \hat{x}_2) \in X$ it holds that $h(\hat{x}) = d^1$ if and only if $\hat{x}_1 < x^0$. Then we claim that by putting $x_1 = w_1$, $x^* = 1 - x^0$, $f_1(x) = (c+d-1)x + 1 - c$ for all $x \in I$ and $f_2(x) = (a+b-1)x + 1 - a$ for all $x \in I$ the sequences $(u_1, u_2, \ldots)$ and $(x_1, x_2, \ldots)$ generated by Iteration 32 satisfy $x_n = w_n$ for $n = 1, 2, \ldots$ and moreover, $u_n = 0$ if and only if $a_n = d^2$. This claim follows also by induction analogously to the first case above.

In the third case suppose that there exists some $x^0 \in I$ such that for any $\hat{x} = (\hat{x}_1, \hat{x}_2) \in X$ it holds that $h(\hat{x}) = d^2$ if and only if $\hat{x}_1 \leq x^0$. Then we claim that by putting $x_1 = v_1$, $x^* = x^0$, $f_1(x) = (c+d-1)x + 1 - d$ for all $x \in I$ and $f_2(x) = (a+b-1)x + 1 - b$ for all $x \in I$ the sequences $(u_1, u_2, \ldots)$ and $(x_1, x_2, \ldots)$ generated by Iteration 32 satisfy $x_n = v_n$ for $n = 1, 2, \ldots$ and moreover, $u_n = 0$ if and only if $a_n = d^2$. This claim follows also by induction analogously to the first case above.

In the fourth and last case suppose that there exists some $x^0 \in I$ such that for any $\hat{x} = (\hat{x}_1, \hat{x}_2) \in X$ it holds that $h(\hat{x}) = d^2$ if and only if $\hat{x}_1 < x^0$. Then we claim that by putting $x_1 = w_1$, $x^* = 1 - x^0$, $f_1(x) = (a+b-1)x + 1 - a$ for all $x \in I$ and $f_2(x) = (c+d-1)x + 1 - c$ for all $x \in I$ the sequences $(u_1, u_2, \ldots)$ and $(x_1, x_2, \ldots)$ generated by Iteration 32 satisfy $x_n = w_n$ for $n = 1, 2, \ldots$ and moreover, $u_n = 0$ if and only if $a_n = d^1$. This claim follows also by induction analogously to the first case above.

This finishes the proof for the case that h has a threshold structure for state $i = 1$. For the case that h has a threshold structure for state $i = 2$ it could be proved similar as for $i = 1$ by distinguishing four different cases and obtaining the appropriate x_1, x^*, f_1 and f_2 for all these cases. However, it follows more elegant by noting that if $S = \{1, 2\}$ any threshold structure for state $i = 2$ is equivalent to a threshold structure for state $i = 1$ and vice versa. For example suppose there exists some $x^0 \in I$ such that for any $\hat{x} = (\hat{x}_1, \hat{x}_2) \in X$ it holds that $h(\hat{x}) = d^1$ if and only if $\hat{x}_2 \leq x^0$. This is a threshold structure according to Definition 35 for state $i = 2$. Obviously it is equivalent to $h(\hat{x}) = d^2$ if and only if $\hat{x}_1 < 1 - x^0$ which gives a threshold structure according to Definition 35 for state $i = 1$. ■

Theorem 38 is a main result in this paper which is based on combining Theorem 23, Theorem 14, Corollary 15, Proposition 29, Proposition 37 and Theorem 33.

Theorem 38 *Consider a $\mathcal{D} = \{d^1, d^2\}$ restricted MDP with state space $S = \{1, 2\}$. Let $a, b, c, d \in I$ be such that (35) holds where P is the transition matrix corresponding to d^1 and Q is the transition matrix corresponding to d^2. Suppose that $1 \leq a + b < 2$, $1 \leq c + d < 2$ and Condition 36 is satisfied. Then for the equivalent full observation MDP there exists some optimal stationary deterministic Markovian policy $\widetilde{\pi}$ having a threshold structure as in Definition 35.*
Moreover, let $\widetilde{\omega} = (y_1, a_1, y_2, a_2, \ldots) \in \widetilde{\Omega}$ be an associated sample path for $\widetilde{\pi}$ and let $f_1, f_2 : I \to I$ be linear functions as obtained in the proof of Proposition 37. If there exists some $x \in I$ for which $f_1(f_2(x)) \geq f_2(f_1(x))$ then for the $\mathcal{D} = \{d^1, d^2\}$ restricted MDP there exist an optimal $\mathcal{D}$-mixing policy which is deterministic and the corresponding decision sequence of zeros and ones is a regular sequence. In particular there exist some $n \in \mathbb{N}$ such that for

all positive integers $t \geq n$ the infinite sequence of decision rules $(a_t, a_{t+1}, \ldots)$ determines an optimal $\mathcal{D}$-mixing policy for which the corresponding sequence of zeros and ones is a regular sequence.

Proof. We have $P = \begin{pmatrix} a & 1-a \\ 1-b & b \end{pmatrix}$ and thus by (19) it is easily seen that $\rho_0(P) = |a+b-1|$. Since $1 \leq a+b < 2$ it follows that $\rho_0(P) < 1$ and similarly we also have that $\rho_0(Q) < 1$. Thus Theorem 23 is applicable for $N = 1$ and thus it follows for the equivalent full observation MDP that Condition 17, Condition 13 and Condition 12 are satisfied. Thus there exist optimal stationary deterministic Markovian policies for the full observation MDP and since Condition 36 is also satisfied it follows that there exists some optimal stationary deterministic Markovian policy $\widetilde{\pi}$ having a threshold structure as in Definition 35.

Let $\widetilde{\omega} = (y_1, a_1, y_2, a_2, \ldots) \in \widetilde{\Omega}$ be an associated sample path for $\widetilde{\pi}$ as in Proposition 37. For the $\mathcal{D} = \{d^1, d^2\}$ restricted MDP we have by Theorem 14 and Corollary 15 that all deterministic $\mathcal{D}$-mixing policies π_t, $t = 1, 2, \ldots$ given by the infinite sequence of decision rules $(a_t, a_{t+1}, \ldots)$ have the same performance. Moreover, this performance is optimal with respect to maximizing the long-run average reward for the $\mathcal{D}$ restricted MDP since $(y_1, a_1, y_2, a_2, \ldots)$ is a sample path of policy $\widetilde{\pi}$ which is optimal for the equivalent full observation MDP. Thus for all $t = 1, 2, \ldots$ policy π_t is an optimal $\mathcal{D}$-mixing policy.

Let $U := (u_1, u_2, \ldots)$ and $(x_1, x_2, \ldots)$ be the infinite sequences generated by Iteration 32 for linear functions f_1, f_2 and appropriate $x_1, x^* \in I$ as in Proposition 37. Then U is an infinite sequence of zeros and ones and by Proposition 37 we have that either (33) or (34) holds. Moreover, according to the proof of Proposition 37 we may assume that either the slope of f_1 is $a+b-1$ and the slope of f_2 is $c+d-1$ or the slope of f_1 is $c+d-1$ and the slope of f_2 is $a+b-1$. Anyway it follows that f_1 and f_2 are monotonically increasing functions since $a+b \geq 1$ and $c+d \geq 1$. Moreover, it follows that the composite functions $f_1 \circ f_2$ and $f_2 \circ f_1$ are both linear functions with slope $(a+b-1)(c+d-1)$ mapping I to I. Hence $f_1(f_2(x)) \geq f_2(f_1(x))$ for some $x \in I$ implies that $f_1(f_2(x)) \geq f_2(f_1(x))$ for all $x \in I$. Thus if $f_1(f_2(x)) \geq f_2(f_1(x))$ for some $x \in I$ then the properties demanded in Theorem 33 for the functions f_1 and f_2 are satisfied and thus the sequence U generated by Iteration 32 is an eventually regular sequence. Thus there exists some $n \in \mathbb{N}$ such that for every positive integer $t \geq n$ the infinite sequence $U_t := (u_t, u_{t+1}, \ldots)$ is a regular sequence of zeros and ones.

Recall from Section 4 that by convention symbol 1 corresponds to action d^1 and symbol 0 corresponds to action d^2. Following this convention let $U' := (u'_1, u'_2, \ldots)$ be the infinite sequence of zeros and ones corresponding to $(a_1, a_2, \ldots)$. Then for $t = 1, 2, \ldots$ we have that $U'_t = (u'_t, u'_{t+1}, \ldots)$ is the infinite sequence of zeros and ones corresponding to optimal $\mathcal{D}$-mixing policy $\pi_t = (a_t, a_{t+1}, \ldots)$. In case (33) holds then it follows that $u'_n = 1 - u_n$ for $n = 1, 2, \ldots$. Thus by property 4 of Proposition 29 it follows for $t = 1, 2, \ldots$ that U'_t is regular of density $1 - \theta$ if and only if U_t is regular of density θ. Thus for every positive integer $t \geq n$ the sequence U'_t corresponding to optimal $\mathcal{D}$-mixing policy π_t is regular since

U_t is regular for $t = n, n+1, \ldots$. In the other case that (34) holds then it follows that $u'_n = u_n$ for $n = 1, 2, \ldots$. Thus it follows for $t = 1, 2, \ldots$ that sequence U'_t is exactly the same as sequence U_t. Thus for every positive integer $t \geq n$ the sequence U'_t corresponding to optimal $\mathcal{D}$-mixing policy π_t is regular of some density θ since U_t is regular of some density θ for $t = n, n+1, \ldots$. ■

In Example 39 we apply Theorem 38 to the $\mathcal{D} = \{d^1, d^2\}$ restricted MDP of Example 7.

Example 39 Consider once again the $\mathcal{D} = \{d^1, d^2\}$ restricted MDP of Example 7. Let $a, b, c, d \in I$ be defined as in Theorem 38. For this example we have that $a = 1$, $b = 0.8$, $c = 0.7$ and $d = 1$. Thus the conditions $1 \leq a + b < 2$ and $1 \leq c + d < 2$ are satisfied. Moreover, recall from Example 22 that Condition 17, Condition 13 and Condition 12 are satisfied. Thus for the equivalent full information MDP there exist some optimal stationary deterministic Markovian policy $\widetilde{\pi}$. We will not prove that also Condition 36 is satisfied, but we note that it seems plausible. Indeed let p be the probability that the machine is in state 1 (the bad state) at a decision epoch. Indeed it seems plausible that there exists some threshold probability p^* such that if p is smaller than p^* that then it is optimal to choose action 1 (work), while if p is larger than p^* that then it is optimal to choose action 2 (repair). Thus assume Condition 36 is satisfied and that policy $\widetilde{\pi}$ has indeed a threshold structure. Let $\widetilde{\omega} = (y_1, a_1, y_2, a_2, \ldots) \in \widetilde{\Omega}$ be an associated sample path for $\widetilde{\pi}$ and for $n = 1, 2, \ldots$ let $v_n, w_n \in I$ be such that $y_n = (v_n, w_n)$ for $n = 1, 2, \ldots$ as in Theorem 38.

Now we distinguish two cases of plausible threshold structures that optimal policy $\widetilde{\pi}$ could have in this example. In the first case suppose there exists some $p^* \in I$ such that policy $\widetilde{\pi}$ chooses decision rule d^1 and thus action 1 (work) if and only if $p \leq p^*$ and chooses decision rule d^2 and thus action 2 (repair) if and only if $p > p^*$. Then following the proof of Proposition 37 we put $x_1 = v_1$, $x^* = p^*$, $f_1(x) = (a + b - 1)x + 1 - b = 0.8x + 0.2$ and $f_2(x) = (c+d-1)x+1-d = 0.7x$. Then $f_1(f_2(x)) = 0.56x+0.2$ and $f_2(f_1(x)) = 0.56x+0.14$. Thus for this threshold structure (32) is indeed satisfied.
In the second case suppose there exists some $p^* \in I$ such that policy $\widetilde{\pi}$ chooses at a decision epoch decision rule d^1 and thus action 1 (work) if and only if $p < p^*$ and chooses decision rule d^2 and thus action 2 (repair) if and only if $p \geq p^*$. Then following the proof of Proposition 37 we put $x_1 = w_1$, $x^* = 1 - p^*$, $f_1(x) = (c + d - 1)x + 1 - c = 0.7x + 0.3$ and $f_2(x) = (a+b-1)x+1-a = 0.8x$. Then $f_1(f_2(x)) = 0.56x+0.3$ and $f_2(f_1(x)) = 0.56x+0.24$. Thus also for this threshold structure (32) is indeed satisfied.

Thus we may conclude that if Condition 36 is satisfied for this example that then (32) holds for all plausible threshold structures. Then it follows by Theorem 38 that there exists some $n \in \mathbb{N}$ such that for all positive integers $t \geq n$ the infinite sequence of decision rules $(a_t, a_{t+1}, \ldots)$ determines an optimal $\mathcal{D}$-mixing policy for which the corresponding sequence of zeros and ones is a regular sequence. In other words under the assumption that Condition 36 holds it follows for this example that an optimal $\mathcal{D}$-mixing policy is among the deterministic Markovian $\mathcal{D}$-mixing policies for which the corresponding decision sequence is in the set $\mathcal{R}$ of regular sequences of zeros and ones and the maximal performance is obtained by

maximizing performances over $\mathcal{R}$.
Recall from subsection 6.2 that for the $\mathcal{D} = \{d^1, d^2\}$ restricted MDP of Example 7 the maximal performance over $\mathcal{R} \cap W_p(200)$ equals 0.3435 (rounded to 4 decimals) and is obtained by the regular sequence with period cycle $(1, 1, 0, 1, 0, 1, 0, 1, 0)$ of density $\theta = \frac{5}{9}$. Moreover, Figure 1 and Figure 2 did give additional visual support for density $\frac{5}{9}$ maximizing the performance over $\mathcal{R}$. If $\theta = \frac{5}{9}$ indeed maximizes $h(\theta)$ and Condition 36 holds then it follows from Theorem 38 that for this $\mathcal{D} = \{d^1, d^2\}$ restricted MDP the $\mathcal{D} = \{d^1, d^2\}$-mixing policy corresponding to period cycle $(1, 1, 0, 1, 0, 1, 0, 1, 0)$ (with symbol 1 corresponding to choosing d^1 and symbol 0 corresponding to choosing d^2) is optimal and the maximal long-run average reward is 0.3435 which is obtained by this policy.

Vice versa this would imply that there should exist some p^* and corresponding threshold property as described above such that for the equivalent full observation MDP and the corresponding policy $\widetilde{\pi}$ induces for any initial state distribution $x \in X$ an infinite sequence of decision rules which is eventually periodic according to the period cycle $(1, 1, 0, 1, 0, 1, 0, 1, 0)$. Indeed this is the case for $p^* = 0.47$ and in fact also for other values of p^* which are (sufficiently) close to 0.47. The reader may check that by putting $x^* = 1 - p^* = 0.53$, $f_1(x) = 0.7x + 0.3$ and $f_2(x) = 0.8x$ as in the second distinguished case above that then for any $x_1 \in [0, 1]$ the infinite sequence $(u_1, u_2, \ldots)$ of zeros and ones obtained according to Iteration 32 eventually becomes periodic with period cycle $(1, 1, 0, 1, 0, 1, 0, 1, 0)$. Moreover, for initial distribution $y_1 = (1 - x_1, x_1) \in X$ the sample path $\widetilde{\omega} = (y_1, a_1, y_2, a_2, \ldots) \in \widetilde{\Omega}$ obtained by applying the threshold property $a_n = d^1$ if and only if $y_n \cdot (1, 0) < 0.47$ satisfies (34). Thus for this $\mathcal{D}$ restricted MDP we have estasblished some additional confirmation for the optimality of the $\mathcal{D} = \{d^1, d^2\}$-mixing policy corresponding to a regular decision sequence with period cycle of density $\frac{5}{9}$ yielding a performance of 0.3435.

7.1 Concluding remarks

In this final section we have shown that for a class of $\mathcal{D}$ restricted MDP the optimality of a deterministic policy corresponding to a regular sequence is assured if some threshold condition is satisfied for the corresponding full observation MDP. In the present paper we have not investigated whether the threshold condition actually holds for the corresponding full observation MDP. However, for many comparable MDP such threshold structure of optimal stationary policies has been investigated and established. For example in section 5.3 of [22] for the so-called searching for a moving target problem it was conjectured that the optimal policy has a simple threshold structure. Namely search location 1 if and only if at the decsion epoch the probability p that the target is at location 1 is larger (or equal) than a certain threshold probability p^*. In [18] the existence of such optimal threshold probability p^* and corresponding policy is proved for many cases of such searching for moving target MDP. Condition 36 for MDP associated with $\mathcal{D} = \{d^1, d^2\}$ restricted MDP is similar and possibly for some problem classes it can be established by similar methods as in [18].

If Condition 36 indeed holds then the (desired) optimality within the class of policies cor-

responding to regular sequences follows if some additional (and easy-checkable) technical conditions (see Theorem 38) are satisfied for the transition matrices induced by the applicable decision rules in $\mathcal{D}$. Note that we have proved that these additional conditions stated in Theorem 38 are sufficient, but possibly these technical conditions can be weakened. Moreover, Theorem 38 can possibly be generalized to $\mathcal{D} = \{d^1, d^2\}$ restricted MDP with S consisting of more than 2 states. Another interesting issue is whether the results on optimality of regular sequences can be extended from the relatively simple threshold structure given by Condition 36 to more involved cases where for example an optimal stationary policy is determined by multiple thresholds.

References

[1] B. Altman, E. Gaujal and A. Hordijk. *Discrete-Event Control of Stochastic Networks: Multimodularity and Regularity.* Lecture Notes in Mathematics. Springer Verlag, 2003.

[2] E. Altman, B. Gaujal, and A. Hordijk. Balanced sequences and optimal routing. *Journal of the ACM*, 47:752–775, 2000.

[3] E. Altman, B. Gaujal, and A. Hordijk. Multimodularity, convexity and optimization properties. *Mathematics of Operations Research*, 25:324–347, 2000.

[4] E. Altman, B. Gaujal, A. Hordijk, and G. Koole. Optimal admission, routing and service assignment control: the case of single buffer queues. In *the 37th IEEE Conference on Decision and Control*, volume 2, pages 2119–2124, Tampa, FL, USA, 1998.

[5] S. Bhulai, T. Farenhorst-Yuan, B. Heidergott, and D.A. van der Laan. Optimal balanced control for call centers. Technical report, Tinbergen Institute, 2010.

[6] E. Fernández-Gaucherand, A. Araposthathis, and S.I. Marcus. On the average cost optimality equation and the structure of optimal policies for partially observable markov decision processes. *Annals of Operations Research*, 29:439–470, 1991.

[7] E. Fernández-Gaucherand, A. Araposthathis, and S.I. Marcus. Remarks on the existence of solutions to the average cost optimality equation in markov decision processes. *Systems and Control Letters*, 15:425–432, 1991.

[8] B. Gaujal, A. Hordijk, and D.A. van der Laan. On the optimal policy for deterministic and exponential polling systems. *Probability in the Engineering and Informational Sciences*, 21:157–187, 2007.

[9] B. Hajek. Extremal splittings of point processes. *Mathematics of Operations Research*, 10(4), 1985.

[10] B. Heidergott and A. Hordijk. Taylor series expansions for stationary markov chains. *Advances in Applied Probability*, 35:1046–1070, 2003.

[11] B. Heidergott, A. Hordijk, and M. van Uitert. Series expansions for finite-state markov chains. *Probability in the Engineering and Informational Sciences*, 21:381–400, 2007.

[12] B. Heidergott and F. Vázquez-Abad. Measure valued differentiation for markov chains. *Journal of Optimization and Applications*, 136:187–209, 2008.

[13] B. Heidergott, F. Vázquez-Abad, G. Pflug, and T. Farenhorst-Yuan. Gradient estimation for discrete-event systems by measure-valued differentiation. *Transactions on Modeling and Computer Simulation*. Accepted.

[14] O. Hernández-Lerma and J.B. Lasserre. *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer, 1996.

[15] A. Hordijk and D.A. van der Laan. On the average waiting time for regular routing to deterministic queues. *Mathematics of Operations Research*, 30:521–544, 2005.

[16] G. Koole. On the static assignment to parallel servers. *IEEE Transactions on Automatic Control*, 44:1588–1592, 1999.

[17] M. Lothaire. *Algebraic Combinatorics on Words*. Cambridge University Press, 2002.

[18] I.M. MacPhee and B.P. Jordan. Optimal search for a moving target. *Probability in the Engineering and Informational Sciences*, 9:159–182, 1995.

[19] M. Morse and G.A. Hedlund. Symbolic dynamics ii - sturmian trajectories. *American Jornal of Mathematics*, 62:1–42, 1940.

[20] G. C. Pflug. *Optimization of Stochastic Models*. Kluwer Academic Publishers, 1996.

[21] M. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley and Sons, 1994.

[22] S.M. Ross. *Introduction to Stochastic Dynamic Programming*. Academic Press, 1983.

[23] R. Tijdeman. Fraenkel's conjecture for six sequences. *Discrete Mathematics*, 222:223–234, 2000.